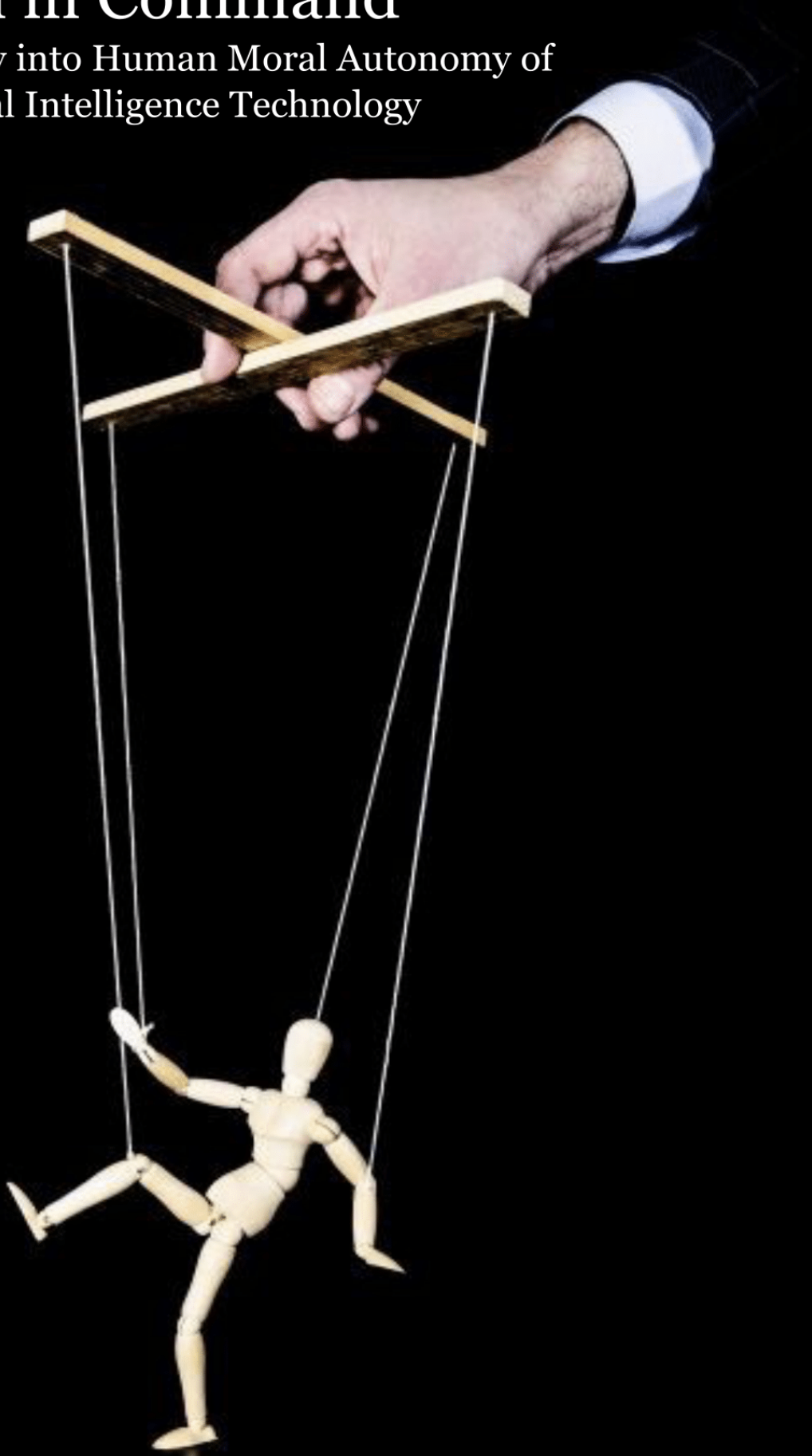
The Human in Command

An exploratory study into Human Moral Autonomy of Behavioural Artificial Intelligence Technology



Can E. Yildiz
4910508
MSc Engineering and Policy Analysis

This page was intentionally left blank

August 26, 2021
Faculty of Technology, Policy and Management
Delft University of Technology

The Human in Command

An exploratory study into human moral autonomy of Behavioural Artificial Intelligence Technology

Thesis submitted in fulfilment of the requirements for the degree of Master of Science in Engineering Policy Analysis

by

Can E. Yildiz
Student number: 4910508

Graduation committee

Chair/First supervisor:	Prof.dr.ir. I.R. van de Poel, Ethics/Philosophy of Technology
Second Supervisor:	Dr. L.J. (Rens) Kortmann, Multi-Actor Systems
External Supervisor:	Prof.dr.ir. C.G. Chorus, Councyl

To be defended in public on September 9^{th} 2021

TU Delft Delft University of Technology

COUNCYL BEHAVIORAL AI TECHNOLOGIES

Acknowledgements

The impact of emerging technologies is often discussed in the present public debate with cynical and optimistic actors on opposite sides. Not without reason, as both tragedies and successes appear where human and computers interact with one another. However, I did not expect to get involved in such topic like human moral autonomy before my thesis preparation. It turned out to be one of the most exhilarating and challenging periods of my life as yet. I spoke with various professors and postdocs in the field of ethics and technology, often resulting in endless conversations on how to design morally responsible technologies. Reading countless papers and books on AI ethics made me often feel drowning in the amount and complexity of information. Thus, without the support of others I would not be able to deliver this end product.

Councyl, the TU Delft start-up company granted me this amazing opportunity to study an emerging technology. The great team of people always encouraged me to utilise my full potential. Thank you, Nicolaas, Hubert and Annebel, for the enjoyable workdays and meetings we had on Fridays. The company radiates refreshing energy because of all of you, and I am sure you will succeed in your ambitions.

I would also like to thank my graduation committee from the TU Delft. Thank you Ibo for guiding me through the amount of complex ethical theories. Every time I got stuck in moral (and often vague) concepts, the meetings with you resulted in clear outcomes. Your expertise in the field of ethics and technology appeared indispensable during my research. A special thanks to Rens, who always asked the most valuable questions and easily noticed potential pitfalls. You frequently expressed your authentic point of view, which induced interesting discussions throughout our meetings. More importantly, you often summarised highly complex matters in just a couple of sentences. I consider that to be a gift. Caspar, from the beginning we met I was inspired by your knowledge and enthusiasm. You possess the capability of energizing every single person around you. During our meetings you usually dragged me out of the details and illuminated an overview of the entire research. You remain a true example to me; not only as a professor or mentor, but also as a human.

Furthermore, I would like to thank family and friends for their unconditional support and love. Zeki, Fadima, Yaren, Ferhat and all other family members, thank you for believing in me and always supporting my ambitions. Thank you Megan for your boundless love and support at all times. Thank you to the most amazing friends Jesper, Thijmen, Luuk, Gilbert, The4ce and all others. And finally, I would like to thank my lovely friends whom I studied with a countless number of times. Boris, Esmée and Roos, our study gatherings were highly motivational and exciting.

Lastly, I would like to mention my deceased grandfather Suleyman from Dersîm, Kurdistan, whom I have never met. Your picture has been in my room for years. Belonging to a family that bears the history of "No friends but the mountains", your photo reminded me of my descent with pride.

Can Eren Yildiz
26 August, 2021
Delft, The Netherlands

Executive summary

The accelerating development of algorithms causes a disruptive effect in many domains, including the complex decision-making of knowledge workers. Experts can manage difficult but repetitive decisions with software technologies like a Decision Support System (DSS). A DSS is used to monitor decisions, get additional insights and improve decisions over time. Their supportive performance characterises these systems to assist human decision-makers. To answer to the pressing demand for transparency in DSSs, Councyl developed Behavioural Artificial Intelligence Tenchnology (BAIT). BAIT is a DSS that adequately supports experts with making decisions. However, algorithms like BAIT may affect the autonomy of experts and their decisions in numerous ways. This thesis studies the human moral autonomy (HMA) of end-users in the context of BAIT. We do this by measuring perceptions of end-users. The product arising from this study is the HMA Survey. On that account, we aim to answer the following research question: "*How can we define, operationalise and measure the (perceived) human moral autonomy of decision support systems like behavioural artificial intelligence technology?*"

The main research question translates into the five following sub-questions:

1. In what ways does BAIT support human decision-making in organisations, and how does that affect task allocations between the human and the DSS?
2. What is a philosophical definition of human moral autonomy in the context of a DSS like BAIT?
3. How can we measure the degree to which DSSs like BAIT respect human moral autonomy using a survey?
4. To what extent does BAIT respect the conditions for HMA according to end-users?
   (a) What importance do end-users ascribe to the philosophical conditions for HMA?
   (b) Does BAIT respect the philosophical conditions for HMA in the perception of end-users?
5. How can DSSs like BAIT be designed so that they better respect human moral autonomy?

BAIT is mathematically founded on Discrete Choice Modelling (DCM). BAIT elicits the preferences of individuals on the decision-making and presents the domain knowledge in quantitative variables. The codification of expert knowledge is an instrument to communicate the advice to the end-user. This knowledge is obtained from stated choice experiments, in which the trade-offs between criteria are weighed with each decision. BAIT guides the end-user by presenting the relative importance between criteria. Moreover, it expresses a percentage that indicates the share of respondents - who participated with the choice experiment - would decide to act upon a decision. Based on these features, BAIT can at least suggest one alternative to the end-user. Additionally, BAIT can be advanced to independently execute tasks, by which the end-user is only expected to supervise the model. We refer to this as 'partial agent autonomy': the system manages the tasks, whereas the human solely evaluates the decisions. Depending on the applied automation level, BAIT decreases the autonomy of end-users in varied extent.

We studied the definition of HMA and the way it manifests within digital environments. Accordingly, we philosophically defined HMA in the context of BAIT as: *A person is morally autonomous if and only if he bears the responsibility and authority for moral supervision on the situation and the decision support system.* Subsequently, we defined conditions under which we consider humans to be morally autonomous in the decision-making with BAIT. This definition resulted in four conditions: (1) inscrutinizability, (2) pressure condition, (3) error condition and (4) critical questioning

and evaluation. The literature argues those conditions to be decisive in determining the autonomy of humans in digital decision-making. The conditions are sub-divided into twelve sub-conditions; collectively, they embody the HMA framework. Hence, we characterised the concept of HMA and, more specifically, operationalised them by setting up (sub-)conditions.

We selected five out of twelve sub-conditions to set up the HMA survey. We added the HMA construct to directly measure the perceived autonomy of users and evaluate the relation with other constructs. Hence, we operationalised the following six theoretical constructs: (1) HMA, (2) accessibility, (3) tractability, (4) getting high-fidelity human expertise, (5) identify & empower human expertise and (6) critical questioning. HMA is the overarching construct, of which the latter five are the sub-conditions. The methodology of the study entails a mixed-method approach. The method contains two components: a pilot survey (i.e. interviews) and the HMA survey. By doing so, we could interpret the results of all parts coherently. The pilot survey fulfils the role of a preliminary validation. We constructed the survey as such to perform factor analysis. This enabled us to validate the hypothesised constructs we defined earlier in the HMA framework. Additionally, the survey results enable to serve a descriptive study to find exciting relationships between constructs according to the perceptions. To conclude, we were able to set up a measurement instrument - that implicitly reflects the hypothesised constructs - by designing a survey.

Three organisations participated in our study: one consulting company and two hospitals. The interviews resulted in insightful responses to improve the survey. Moreover, the respondents proved to be sufficiently familiar with the technology and the underlying concepts of the survey. The survey contains two types of statements: perception statements and importance statements. The former measures the perception of a construct related to BAIT, whereas the latter measures the importance of each construct. The number of respondents are too low to validate the correlation structure of the survey, but we were able to form a new set of constructs. The new constructs in the validated HMA survey are (1) Choice reflection, (2) Autonomy, (3) Boundless decision-making, (4) Intelligibility, (5) Explainability (6) Knowledge interchange. The validated version contains statements of the overarching HMA, critical questioning and accessibility constructs primarily. Hence, it contains disproportionally more statements of those constructs than the remaining constructs.

Additionally, we conducted descriptive analysis and formulated guidelines to improve the technology in terms of HMA. We found discrepancies between the perception and importance statements. The mean difference between both types of statements are validated on significance for each construct. By doing so, we were able to identify the disparity between the perception of the respondents and the importance they ascribe to each construct. End-users predominantly scored positive on each of the perception statements, indicating they consider themselves (relatively) autonomous in using BAIT. However, the importance statements indicated even a higher score. Therefore, we recommend the problem owner to use the divergence to improve BAIT. We also found in our analysis a positive relationship between HMA and all other sub-conditions. Although we could not draw conclusive evidence from the data, the current findings show confirmatory patterns in relation to the underlying concepts. After analysing the data and conducting additional literature review, we formulated recommendations to improve the technology in terms of the above-mentioned constructs. This supports the problem owner, Councyl, to improve BAIT suitably to adhere to the conditions of HMA.

List of Figures

List of Tables

List of Acronyms

Table 1: List of acronyms

Acronym	**Meaning in thesis**
AA	Artificial agent
AI	Artificial intelligence
ADM	Automated decision-making
ANN	Artificial neural networks
BAIT	Behavioural artificial intelligence technology
DCM	Discrete choice model
DSS	Decision support system
EA	Ethical acceptability
ES	Expert system
HA	Human agent
HCI	Human-computer interaction
HMA	Human moral autonomy
IC	Intensive care
LOA	Level of automation
MNL	Multinomial logit
NEC	Neonate with pre–necrotizing enterocolitis
OLVG	Onze Lieve Vrouwe Gasthuis (Hospital)
RRM	Random regret minimization
SA	Social acceptance
TDP	Team design pattern
UMCG	Universitair Medisch Centrum Groningen (Hospital)

Contents

Chapter 1

Introduction

> Where does the good lie? 'In choice.' Where does the bad lie? 'In choice.' And that which is neither good nor bad? 'In things that lie outside the sphere of choice'
>
> Epictetus, Discourses 2.16

"Computer says no" is a well-known phrase duped parents read in the automatically generated letters they received. Whenever they asked for explanations, the government officials referred to the decision made by the system and classified the parents as fraudulent behind the scenes. This narrative exemplifies the procedures in the Dutch childcare benefits scandal, in which decision-makers did not acknowledge the responsibility and consequences of the decision they made. This is because a vast system that the officers did not understand to the full extent gave output they were expected to comply. Government officers stated they complied with the system's decisions and obeyed the law to their ability, by which they concluded they could not and should not be held accountable afterwards. From this, we can conclude an excess reliance of decision-makers on digital systems may have catastrophic consequences. Hence, the Dutch childcare benefits scandal exemplifies the importance of decision-makers' Human Moral Autonomy (HMA). This is not only to held decision-makers accountable for the consequences of made decisions but also to ensure all cogs in the bureaucratic machine contribute to preventing such scandals ever again.

The aggregated amount of data has been increasing exponentially in the last couple of decades, enhancing the evolving number of artificial intelligence (AI) models to enable automated decision-making (ADM) tools at a societal level (Araujo, Helberger, Kruikemeier, & de Vreese, 2020). This advancement caused an ongoing public debate in society to move beyond traditional AI algorithms and integrate ethical and legal principles in such models to preserve human rights in the new digital era (Ntoutsi et al., 2020). Predominantly, the ethical concerns are related to the interaction between big data and machine learning models, mostly being perceived as black boxes in the form of highly complex neural networks (Adadi & Berrada, 2018). A legitimate integration of such models demands transparent and explainable (XAI) perspectives to enhance credibility from society towards them (Tjoa & Guan, 2020).

The foundations of Discrete Choice Models (DCM) used at Councyl[1], the problem owner, fundamentally differs from both machine learning's neural networks approach and conventional business rules. Councyl claims that BAIT does not need explicit knowledge as input, performs no black/white decision structure and is not as rigid as business rules are. Moreover, Councyl declares that BAIT does not need high-quality big data, is relatively more transparent, and possesses better features to identify bias compared to machine learning. On the other hand, Councyl mentions BAIT does represent some of the beneficial components of both concepts. It is similarly transparent and as pragmatic as business rules but does not give in on flexibility and nuance as characterised by machine learning. Given these claims, Councyl may fill in the market gap regarding the digital transition of decision-making problems caused by a lack of focus on moral autonomy. This potential market gap provides a research opportunity to study the degree of HMA from a normative and

[1]Councyl is a start-up company that employs interpretable decision support systems based on discrete choice models for automation purposes. For more info: www.councyl.ai

descriptive point of view.

The objective of DCMs is to predict choice outcomes among a given set of discrete alternatives to identify the behaviour of each respondent based on the decision rule (Sifringer, Lurkin, & Alahi, 2020). Decision rules are the structure humans use when making choices (Payne, Payne, Bettman, & Johnson, 1993). The choice modelling domain is substantially built upon the Multinomial Logit Model, which contains a simple utility specification for analysing discrete choices. Dietvorst, Simmons, and Massey (2015) state the algorithm type plays a dominant role in the extent of aversion from society against them. At Councyl, they appear as decision support systems (DSS): interactive, computer-based information systems that utilise decision rules and models, coupled with a comprehensive database (Turban & Watkins, 1986). These knowledge-based systems (or expert systems) contain codified human knowledge used in situations that generally require human intelligence to solve complex problems (Abraham, 2005). This specification is captured in the term Behavioural Artificial intelligence Technology (BAIT) (ten Broeke, 2020). The fundamental difference between the two distinctively described models enacts potential research opportunities to pursue explainable models within the government.

As for any state-of-the-art technology, it is crucial to understand the consequences of social accountability and responsibility (Cummings, 2006). Taebi (2017) brings two important concepts together from a philosophical point of view to assess philosophical research cases: social acceptance (SA) and ethical acceptability (EA). Social acceptance alludes to the degree of acceptance across concerned communities or society as a whole. In contrast, ethical acceptability implies an ethical reflection on the technology by considering moral matters emerging from the employment. The nature of this study is scientifically referred to as empirical ethics: it entails both normative (ethical), and descriptive (social) approaches to formulate explanations on a phenomenon (Musschenga & Musschenga, 2005). Hence, we aim to provide a description and analysis of the perceptions of decision-makers. On that account, ethical guidelines may support preserving their HMA in the usage of BAIT. The study of Zussman (1992) distinguishes between “how decisions should be made” and “how they are in fact made”. In contrast, this study distinguishes between ”how decisions should be made” (according to HMA literature) and ”how experts perceive their actual HMA” when using BAIT. We create two distinctive lines; the former represents the normative, ethical part, and the latter represents the descriptive, social part. Additionally, respondents will be asked to define their appreciation regarding some of the ethical values.

The particular ethical value that forms the basis of this study is moral autonomy. Moral autonomy is - according to Kant - defined as the combination of freedom and responsibility; it is a submission to laws that one has made for oneself and thus is not subject to the will of another (Wolff, 1998). Scanlon (1972) defines moral autonomy as a state in which a person reckons himself as sovereign in deciding what to believe and in weighing competing reasons for action. Hence, a decrease of the expert’s moral autonomy in the decision-making could lead to an erosion of accountability of the expert (Cummings, 2006).

This chapter presents the research gap in section 1.1, formulates the problem statement in section 1.2, determines the research goal in section 1.3, presents the research questions in section 1.4,determines the final deliverable in section 1.5 and clarifies the research structure in section 1.6.

1.1 Research gap

Within this section a literature study is conducted. We conducted a literature study to moral autonomy, reponsibility, accountability and the societal context. By doing so, we are able to find the research gap. In section 1.1.1 literature is collected and analyzed accordingly. Thereafter, in section 1.1.2 a conclusion is drawn with most the most important findings and the research gap is identified.

1.1.1 Preliminary literature review

The objective of the preliminary literature review is to collect several papers that support identifying the research gap. The key articles are listed in table 1.1. This research is predominantly ethical

but also has interfaces with other research domains. Hence, the papers are categorized based on the focus to group them accordingly.

Table 1.1: Overview of published research with their approach and objectives

Citation	Focus	Title
(Santoni de Sio & van den Hoven, 2018)	Ethics	Meaningful human control over autonomous systems
(Gray, Young, & Waytz, 2012)	Ethics & Psychology	Mind Perception Is the Essence of Morality
(Cummings, 2006)	Ethics & Psychology	Automation and Accountability in Decision Support System Interface Design
(Van den Hoven, 1998)	Ethics	Moral Responsibility, Public Office and Information Technology
(Sternberg, 2012)	Ethics	Model for ethical reasoning
(Dworkin, 2015)	Ethics	The nature of autonomy
(Dworkin, 1981)	Ethics	Moral autonomy
(Waa & Diggelen, 2020)	Ethics & (Psychology)	Allocation of Moral Decision-Making in Human-Agent Teams: A Pattern Approach
(Floridi et al., 2018)	Ethics	AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations
(Floridi, 2019)	Ethics & Society	Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical
(Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016)	Ethics & Society	The ethics of algorithms: Mapping the debate
(Chiodo, 2021)	Ethics	Human autonomy, technological automation (and reverse)
(Nissenbaum, 1996)	Ethics & Society	Accountability in a computerized society

Ethical outlook on HMA

While touching upon many other scientific fields, the papers of Santoni de Sio and van den Hoven (2018), Sternberg (2012), Dworkin (1981), Dworkin (2015) and Chiodo (2021) primarily provide an ethical outlook on HMA. Moral autonomy ethically implies a person is the author or source of their own decisions. Dworkin (1981) exemplifies a fundamental aspect of moral autonomy: the deliberate or undeliberate assignation of something or someone being an epistemological authority. Epistemology is a philosophical domain concerned with knowledge, justification and rationality of belief. Relying on the expertise or judgement of an acknowledged artificial entity leads to a considerable lack of autonomy. Artificial entities within the context of this study are digital systems used by humans for decision-making. The phenomenon of relying upon a system - without being able to raise system independent reasons - is what Van den Hoven (1998) calls 'epistemic enslavement'. Epistemic slaves are only characterised as such when they work in so-called 'narrowly embedded systems'. This phenomenon is often characterised as a situation in which humans lack critical evaluation of the decision-making.

Human-computer interaction outlook on HMA

Cummings (2006), Gray et al. (2012) and Nissenbaum (1996) provide an ethical outlook on human agents in a computerized environment substantiated with psychological consequences. The paper of Cummings (2006) touches upon a couple interesting points. Firstly, she states that a decrease in moral agency likely leads to an erosion of accountability. The ability to harm people through a moral buffer enables decision-makers to distance themselves from their decisions ethically. Moreover, automation bias (i.e. overly trusting information in complex systems) degrades accountability and abandonment of responsibility. The latter can be counteracted by increasing social accountability, i.e. justifying outcomes and decisions by decision-makers. Nissenbaum (1996) conceptually proves the degradation of accountability in a computerized society with disastrous consequences for liability. This is especially the case when we - as a society - give unintentional consent to automated decision-making processes by machines. The paper draws a philosophical account on which four barriers are identified and treated more extensively: (1) the problem of many hands; (2) the problem of bugs; (3) blaming the computer and; (4) software ownership without liability. Accountability is important for many reasons, but philosophically one may state a developed state of responsibility is good in its own right. It can predominantly be reached when moral autonomy is respected in such settings. Psychologically, note that holding people accountable for their actions provides strong motivation to minimize the risk of the consequences.

Societal outlook on HMA

The last categorization in the literature is the societal one. Many ethical frameworks on HMA have been developed on high and lower levels, conveying overlap to some extent (Floridi & Cowls, 2019; Floridi et al., 2018; Mittelstadt et al., 2016). The work of Floridi et al. (2018) provides a high-level framework to create AI for good purposes. Despite focusing on one principle from this framework, it clearly illustrates the importance of each principle and the connection between each one of them. It explains autonomy as the power to decide (to decide) and proposes a concept of 'meta-autonomy': humans should retain power on the decisions that have to be taken. The paper of Mittelstadt et al. (2016) clarifies the importance of public debates on algorithms. Algorithms are challenging because of their complexity, but their uncertainty has vast consequences on a societal level. They mapped ethics of algorithms on various types of evidence, outcomes, and effects. Those types are categorized as epistemic concerns and normative concerns accordingly. Conclusively, this paper explains the main concerns on algorithms and provides concepts of how ethical issues are treated in the literature.

1.1.2 Literature conclusion and research gap

The comprehensive literature review aimed to identify the status quo on the scientific literature of HMA in technological context. Based on the extensive assessment of the literature and literature selection, a number of conclusions can be drawn:

- There is no consensus on the exact definition of HMA and hence specific characterisation is paramount
- However, consensus is identified on the degraded degree of HMA in a computerized environment
- Increased HMA leads to higher social accountability and responsibility
- HMA appears to be crucial in order to maintain the "human touch" in decision-making processes, especially in the public domain
- HMA is a key principle across many state-of-the-art ethical frameworks on AI
- There are no instant ready-to-use frameworks to assess the degree of HMA in DSS

The conclusions mentioned above illuminate the relevance of characterising a satisfactory concept of HMA in DSS and empirically measure the degree of HMA in the usage of such DSS. There is still no ready-to-use measurement instrument to measure perceptions of decision-makers. Helberger, Araujo, and de Vreese (2020) explicitly mentions the need for further research on the perception of ADMs that operate from a hybrid approach, in which humans and DSS co-operate.

The specific ethical principle that will be studied in this research is autonomy, one of the five principles presented in the ethical framework of Floridi et al. (2018). The setting of human-computer interaction of BAIT can be classified as "Supported moral decision making", illustrated in fig. 1.1. The research gap is ethical and covers the lack of knowledge on the degree of HMA of BAIT. The research gap is identified based on two shortcomings. Firstly, the definition and understanding of HMA in the context of BAIT is not studied yet. Second, there is no instrument by which perceptions of HMA can be measured. Hence, the definition of HMA must be defined, operationalised and measured accordingly. By doing so, we enable ourselves to study HMA from both normative and descriptive perspectives.

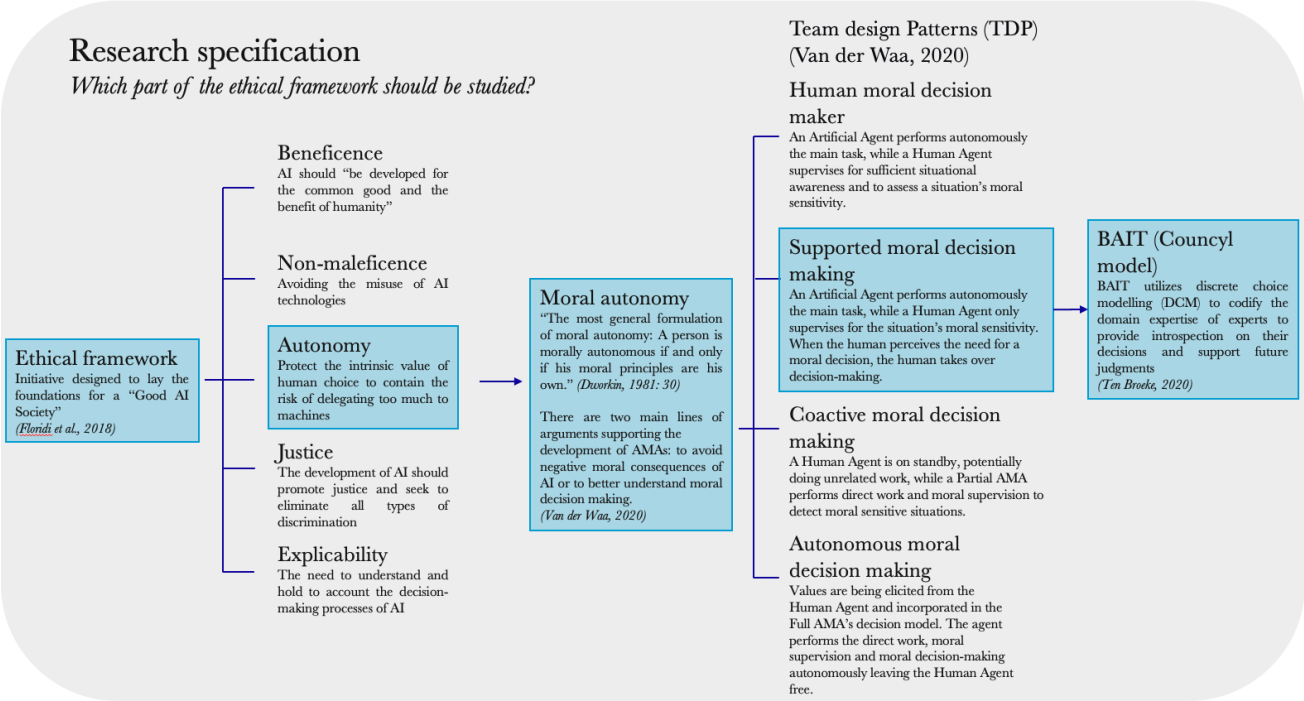


Figure 1.1: Specification of research topic

1.2 Problem statement

Councyl identified with their founding an urgent demand for transparent and interpretable models in the public sector. Some of their potential clients are still concerned with the autonomy of their decision-makers when using such systems. Hence, the problem entails the understanding of moral autonomy and the perception of decision-makers on this phenomenon. The Netherlands Court of Audit explained their latest recommendations on using algorithms within the government in a framework to assess decision-making models. The last perspective - ethics - consists of one paramount value: respecting the moral autonomy of the user. According to Cummings (2006), decision-makers can ethically distance themselves from their actions when there is the ability to harm people through a moral buffer (psychological distancing). The moral buffer within this case is explicitly BAIT through a DSS.

It is essential to understand the degree of HMA of decision-makers' usage of BAIT to prevent ethical distancing. One of the benefits of this research is an increased understanding of the concept of HMA within the described context. By bringing together the normative and descriptive aspects, we enable ourselves to understand the ideally prescribed conditions for HMA by philosophers and comprehend the view of end-users on these conditions. The main benefit for future research is an operationalisation of normative constructs into quantified, measurable variables to measure perceptions of communities. Hence, this induces a step towards a more holistic approach by considering ethical arguments and empirical findings. Consequently, the results of this study serve a for the value-sensitive design of the technology.

1.3 Research goal

The previous sections show a pressing need for a thorough analysis of the HMA of the decision-maker with the use of BAIT. This study aims to reveal the perceptions of decision-makers on their perceived moral autonomy. We do this by defining a suitable measurement instrument on HMA and conducting a survey study to collect data. Altogether, we synthesise the normative and empirical results by (1) defining HMA in the context of BAIT, (2) operationalise normative constructs into empirical variables and (3) measure perceptions of end-users. The creation of a measurement instrument meets this demand. Additionally, we propose recommendations to improve the technology in terms of HMA.

1.4 Research questions

A case study will provide an answer to the following main research question:

Main research question

How can we define, operationalise and measure the (perceived) human moral autonomy of decision support systems like behavioural artificial intelligence technology?

The main research question translates into the following sub-questions:

1. In what ways does BAIT support human decision-making in organisations, and how does that affect task allocations between the human and the DSS?
2. What is a philosophical definition of human moral autonomy in the context of a DSS like BAIT?
3. How can we measure the degree to which DSSs like BAIT respect human moral autonomy using a survey?

Additionally, this research provides the results of the empirical study and guidelines to improve the technology in terms of HMA. These by-products are defined to serve the problem owner, Councyl, in the development of BAIT. Therefore, the following sub-questions are formulated:

4. To what extent does BAIT respect the conditions for HMA according to end-users?
   (a) What importance do end-users ascribe to the philosophical conditions for HMA?
   (b) Does BAIT respect the philosophical conditions for HMA in the perception of end-users?
5. How can DSSs like BAIT be designed so that they better respect human moral autonomy?

1.5 Final deliverable and scope

The final deliverable contains five products. Firstly, we conduct an extensive literature review to conceptualise the meaning and manifestation of HMA in a computerised context. Subsequently, we operationalise the most critical theoretical constructs. We construct a s survey to measure the perceptions of potential users on various conditions of HMA. Altogether, this methodology combines ethical acceptability (EA) and social acceptance (SA). The last two by-products of this research are the results of the empirical study on end users' perceptions and a guideline to improve HMA for BAIT. To conclude, the scope of this research is defined as: "Defining, operationalising and measuring human moral autonomy of users of Behavioural Artificial Intelligence Technology."

1.6 Research approach & structure

This thesis is divided into multiple chapters, each chronologically covering one of the research questions mentioned above. In chapter 2 it is explained what conventional DSSs and ES entail, how BAIT differs in characteristics from those traditional systems and how BAIT is built. This chapter aims to understand how task allocation for BAIT works.

Chapter 3 aims to characterise the HMA of the user and provides additional context to underline the importance of the ethical aspect in advance of BAIT. A literature review is conducted in this part, in which a select number of papers are collected and analysed to extract the essential criteria for HMA. Eventually, this answers the second sub-question by defining a theoretical framework to assess the HMA of the user in the employment of BAIT.

After operationalising the concepts in the theoretical framework, we translate those constructs into a survey to measure them appropriately. The survey is built upon the theory of factor analysis. By doing so, we enable ourselves to measure the perceptions of decision-makers. Additionally, expert interviews will be held to understand the most critical points on HMA as perceived by decision-makers and pretest the survey. After that, we statistically validate the survey utilising factor analysis. This is the first step into creating a valid survey. Collectively, this entails the scientific result of this study.

Whereas the former activities mainly serve the scientific contribution of this study, the following activities serve the interests of the problem owner. We distribute the survey among a small group of decision-makers. After that, we descriptively analyse the results to obtain the perceptions of decision-makers on their perceived HMA. This analysis also enables to demonstrate the relationships between the theoretical constructs, which implicitly embody the survey. Partly based on those findings, we propose guidelines to improve the technology in terms of HMA. We primarily offer guidelines based on best practices from the literature. This will lead to advice that Councyl can use to enhance their DSS to respect the HMA of the decision-makers. Eventually, we discuss the results and conclude that. Lastly, we formulate recommendations for further research. Figure fig. 1.2 illustrates the structure of this research and shows the relation between the multiple parts.

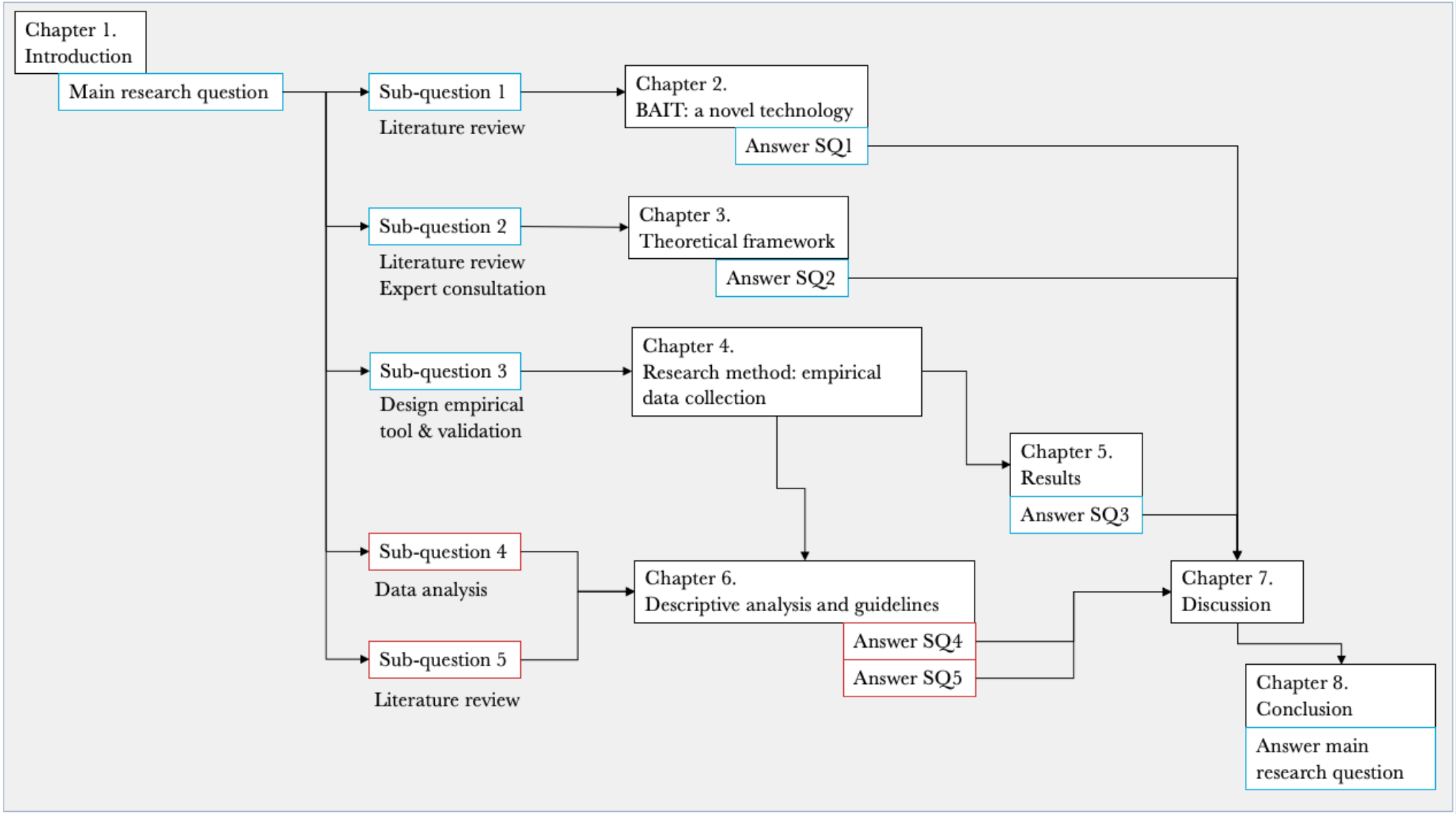


Figure 1.2: Research structure

Summary box chapter 1

The goal of chapter 1 is to present an introductory context, identify the research gap and formulate the research questions. By doing so, we aimed to clarify the state of affairs of Human Moral Autonomy (HMA) in the context of Behavioural Artificial Intelligence Technology (BAIT). A preliminary literature study is conducted to identify the research gap. Thereby, we were able to formulate the problem statement, the research goal and approach. The following points elucidate this chapter:

- Excess reliance of decision-makers on digital systems has troublesome effects on responsibility and human expertise
- BAIT is a novel technology that uses Discrete Choice Modelling (DCM) to provide hypothesised transparent decision support
- This study merely focuses on the HMA of end-users of BAIT
- There are no ready-to-use frameworks to measure HMA of end-users of such technologies accurately
- This study aims to define HMA in the context of BAIT, operationalise and measure end-users' perceptions
- This study is set up to create an HMA survey to measure perceptions of potential end-users
- The study ends with an exciting descriptive analysis of the results and proposes guidelines to improve BAIT in terms of HMA

Chapter 2

BAIT: defining the technology

> "An autonomous person cannot accept without independent consideration the judgment of others as to what he should believe or what he should do. He may rely on the judgment of others, but when he does so he must be prepared to advance independent reasons for thinking their judgment likely to be correct, and to weigh the evidential value of their opinion against contrary evidence."
>
> Scanlon, A Theory of Freedom of Expression, p. 216

This chapter aims to understand the system of BAIT in the context of similar technologies, study the system characteristics, and present the way it supports decisions to experts. Therefore, section 2.1 examines conventional DSSs, section 2.3 compares BAIT with other DSSs and presents the modelling process and section 2.2 entails the mathematical and theoretical foundation of BAIT. Collectively, this chapter formulates an answer to the following sub-question:

Research question 1

In what ways does BAIT support human decision-making in organizations and how does that affect task allocations between the human and the DSS?

2.1 Decision support systems

Because BAIT cannot be considered a rule-based technique or machine learning, it is noteworthy to put the technology in a broader context within this section. Section 2.1.1 explains the manifestation of current DSS and their characteristics. In section 2.3 the comparison i made between conventional DSS applications and BAIT.

2.1.1 Intelligent decision support systems

A simplified theory of decision-making entails two distinctive states in which an individual either prefers to choose one of the alternatives (say A or B) (Edwards, 1954). DSS can be used for monitoring decisions, getting real-time insight, and improving them over time. DSS are interactive, computer-based information systems that utilise decision rules and models, coupled with a comprehensive database (Turban & Watkins, 1986). They mostly appear as rule-based expert systems, which contain human knowledge used in situations that generally require human intelligence to solve complex problems (Abraham, 2005). Preliminary to current development in technology, expert systems were not capable of handling complex functions except for handling purely analytical tasks (Nolan, 1998). With the development of AI in the past 30 years and with the support of statistical techniques, DSSs have become more intuitive and intelligent lately (Weiss & Kulikowski, 1991).

It is paramount to first emphasise the differences between DSS and expert systems (ES) before drawing the conceptual integration between the two concepts. Whereas the objective of DSS is to assist the human decision-maker, for ES, it is mainly to replicate a human advisor (Turban & Watkins, 1986). Moreover, DSS is oriented toward decision-making and ES functions as a transfer of expertise from humans to systems. Lastly, DSS aims to treat unique problems contrary to ES, treating more repetitive decisions. Although both concepts seem to propagate different characteristics, they can be complementary to each other. The interplay between the objectives of AI techniques (e.g. expert systems, among others) and DSS have been conflicting for a long time, as the former aims to replace human tasks and the latter solely aims to support decision tasks (Arnott & Pervan, 2005). One should thereby not overlook the main goals of both concepts: they both seek to improve the quality of decisions (Sen & Biswas, 1985).

The expert system contains three main components that interact with each other. The knowledge base combines and stores the explicit knowledge of numerous experts in different types and rules (Abraham, 2005). The inference engine identifies the problem type and finds the relevant facts or rules from the knowledge base to transfer to the user interface (Nelson Ford, 1985). Hence, it is the crucial component that assembles the correct knowledge based on the interaction between the user and the user interface. The user interface serves the needs of designing, updating and using the expert system. The expert system is shown in fig. 2.1 and clarifies the relationship between the entities as mentioned above.

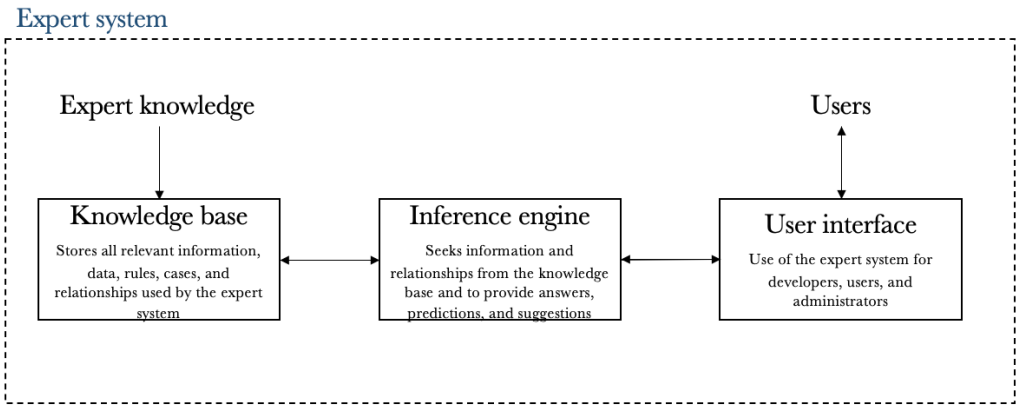


Figure 2.1: Expert system
(Abraham, 2005)

Researchers consider the integration of an ES within a DSS to be an intelligent DSS (IDSS) when it at least partly (and adequately) mimics human intelligence (Guerlain, Brown, & Mastrangelo, 2000). IDSSs predominantly appear within the healthcare sector, providing fascinating insights for clinicians under the title of clinical DSS (Baalen, Boon, & Verhoef, 2021). Although its widespread application in a limited number of industries, it has not always been relatively successful as it is becoming now. Zeleznikow and Nolan (2001) already proved how, employing statistical techniques, they were able to handle a degree of uncertainty. However, any given AI technology can be integrated and displayed through a DSS. This phenomenon is displayed in fig. 2.2

Usually, two main lines of intelligent DSSs predominantly appear in the literature: knowledge-based (e.g. rule-based expert systems) and non-knowledge based systems (e.g. neural networks) (Aronson, Liang, & MacCarthy, 2005). Knowledge-based refers to a set of systems that encode experts' judgements to resemble some of the human intelligence what we call expert systems (Spiegelhalter & Knill-Jones, 1984). Non-knowledge based DSS entails systems using machine learning and other sophisticated statistical techniques to identify patterns (Chung, Boutaba, & Hariri, 2014). Non-knowledge systems are often perceived as 'black boxes': they hardly explain the computation intrinsically in a way that it can be presented to humans (Rudin, 2019). On that account, any explanation for the black box model may draw an unreliable characterization of the original model, with all its consequences for the decision-making (Mittelstadt et al., 2016).

Most knowledge-based systems are characterized by the inability of handling problems for which no rules exist. The self-learning characteristic is paramount in such cases to define new rules to

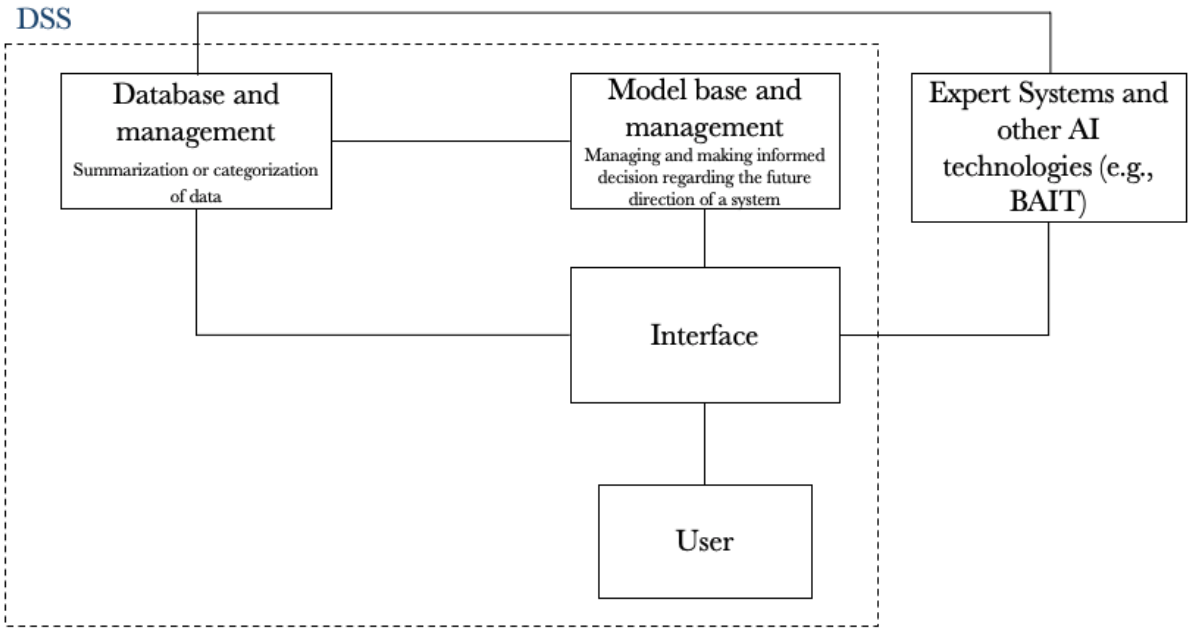


Figure 2.2: Integration of ES in a DSS
(Turban & Watkins, 1986)

tackle future issues. For most expert systems, it is still somewhat a burdensome activity, as there will always remain complex problems for which no rules exist yet (Prentzas & Hatzilygeroudis, 2007). Such techniques, however, have been developed and used more robustly (see Uricchio, Giordano, and Lopez (2004)) to account for uncertainty in decision-making in a complex environment. Additionally, there is no need to train the model on data, enabling the expert to interact instantly. Hence, knowledge-based models are used to deal with primarily repetitive, superficial problems.

Non-knowledge based systems use mainly machine learning techniques to identify patterns independently based on the artificial neural networks (ANN), without any further input of explicit expert knowledge (Marakas, 1999). Maintenance is not performed based on expert knowledge in a straightforward way either. In contrast, most non-knowledge based systems are equipped with self-learning mechanisms to constantly compute the most recent output based on the latest data (Berner & La Lande, 2007). Those mechanisms have, however, implications for the calculation of the model. Data quality is hence paramount, as it will likely provide poor results when the data is either incorrect or even missing (Jordan & Mitchell, 2015). This phenomenon is what we refer to as 'Garbage in is Garbage out'. To conclude, non-knowledge based systems can deal with highly complex problems but generally do this at the cost of interpretability and transparency.

2.1.2 System characteristics of BAIT

The integration between BAIT and DSS essentially defines the terminology of an intelligent DSS based on choice behaviour modelling. Thereby, it aims to combine analytic techniques with conventional data and are characterised by its relatively user-friendly features similar to conventional DSSs (Sprague, 1980). The expert system (BAIT) represents the methodology that performs the internal processes. One could perceive that as an AI, which is fundamentally the driving technology based on which the predictions are made. It is connected to the database and the interface, providing the results and displaying it to the user. The DSS itself contains the interconnected components of a database, model base and interface. This description is shown in fig. 2.2.

The system structure of BAIT is similar to conventional expert systems but has peculiar components that make it a unique technology. It starts with the inventarisation of criteria in which the most critical variables are determined for making specific decisions. Subsequently, a choice experiment is conducted to obtain the relative weights for each of the criteria. This is done to capture the expertise of respondents within the presented context of alternatives. The captured expertise (i.e. like/dislike of criteria) is calculated by decision rules to determine the trade-offs

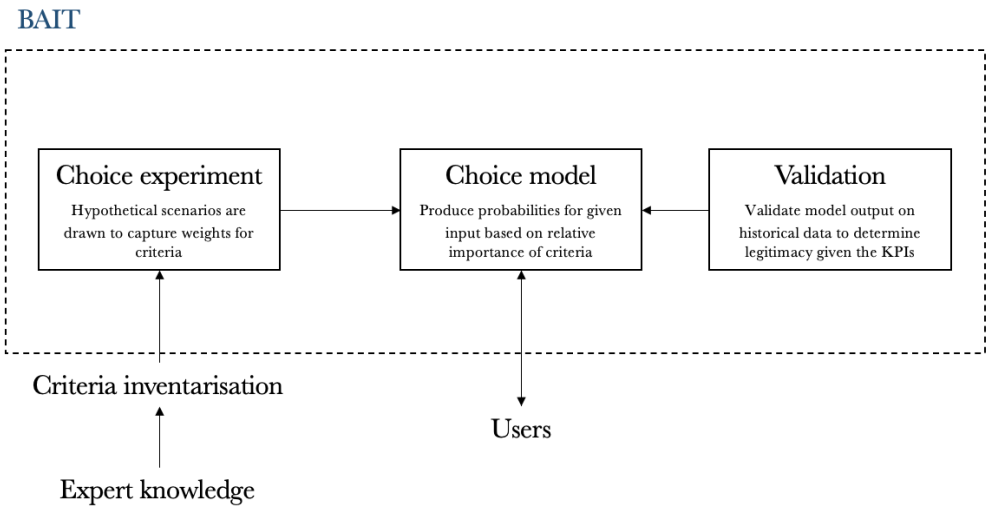


Figure 2.3: BAIT

experts make. The estimated likelihood is calculated for a given outcome to construct the model. In contrast to most simple rule-based systems, BAIT eventually provides an estimated probability indicating the percentage of experts that would advise approving the decision. Consequently, it gives a piece of more nuanced advice rather than the rigorous feature rule-based systems have.

BAIT also has similar characteristics to non-knowledge based systems. Where ANN in machine learning statistically finds patterns in the input data, BAIT merely uses explicitly defined criteria and weights based on the input from the choice experiment. It is for experts rather complicated to understand how machine learning models provide predictions for given decision-making, as it uses highly complex ANN to perform the calculation. This description illustrates the opaque feature of machine learning models and why it is considered to produce black box technology. For BAIT, it is assumed that experts can better understand the reasoning behind the advice; as for each criterion, the weights and input are given. Hence, the technology is more transparent as it provides decision support by displaying the trade-offs between criteria for a particular combination of attribute levels.

2.2 Discrete choice modelling

This section entails the methodology of discrete choice modelling (DCM), the technique behind BAIT that supports codifying domain expertise. section 2.2.1 explains the procedure of choice experiments and section 2.2.2 describes the discrete choice modelling technique and mathematical foundations.

2.2.1 Choice experiments

Choice experiments can be conducted in two distinctive ways: revealed preferences (RP) and stated preferences (SP). RP implicates that choices are proxies for unobserved preferences and hence provide information to model choices in terms of latent utilities (Samuelson, 1948). On the contrary, SP describes potential choices in terms of a set of constructed measures of various combinations of attribute values representing hypothetical alternatives (Hensher & Assoiate, 1993). This research studies the application of BAIT, which almost always is conducted with SP experiments. The main difference between stated and revealed values is referred to as hypothetical bias: alternative specific values primarily differ due to over or underestimation of alternatives by individual respondents (Murphy, Allen, Stevens, & Weatherhead, 2005).

Experimental designs are used to make a limited mapping of the preferences for each subject and estimate multiple measurements that are of interest for a specific choice (M. E. Ben-Akiva, McFadden, Train, et al., 2019, p.8). On that account, experimental designs consist of hypothetical choice scenarios containing various alternatives that respondents have to choose. Thereby, experimental designs reflect real choice scenarios as much as possible to ensure the validity of

data (Molin & Timmermans, 2010). For this research, specifically, the choice scenarios differ from decision-making to another, as every organization aims to optimize theirs, given the challenges they face. Hence, it is crucial to understand the general thread through all choice experiments designed within Councyl.

The hypothetical choice scenarios contain combinations of changing attribute levels to obtain information on the relative importance respondents put on each of the attributes (M. E. Ben-Akiva et al., 2019, p.11). There are two lines of thoughts on the maximum number of attributes. Some researchers argue six is the maximum number of respondents can process. In contrast, others argue inclusion of more attributes is possible as long as consistency and clearance are preserved on the attributes (Hensher, Rose, & Collins, 2011). The choice experiments conducted by Councyl are for pragmatic reasons executed on a small sample of experts. This methodology is theoretically valid as long as the priors for attributes are chosen correctly and results in a smaller sample size requisition given the similar statistical parameter significance (Walker, Wang, Thorhauge, & Ben-Akiva, 2015).

2.2.2 Discrete choice model

The random utility maximization (RUM) is one of the leading mathematical theories of DCM (M. Ben-Akiva & Lerman, 2018). RUM assumes each decision-maker chooses the alternative that generates the highest hypothetical outcome. An outstanding option is random regret minimization (RRM), which implies decision-makers aim to minimize regret when choosing between alternatives (Hensher, Greene, & Chorus, 2013). Although RRM has increased in popularity recently as a complementary modelling paradigm, RUM is the general dominating theory for DCMs (Hensher, Greene, & Ho, 2016).

Utility function

The RUM-MNL model denotes a set of alternatives and aims to calculate the probability that an individual chooses a specific alternative (McFadden, 1986). The utility function consists of a systematic utility (V) and a random utility (ε). Each attribute within the systematic utility function is composed of a parameter (β), which constitutes the weight (i.e. taste) of an attribute, multiplied with the attribute value (x). Hence, the systematic of an alternative contains the sum of all utilities of the attributes for one alternative. The random utility (ε) is a proxy for noise and includes the uncertainty of the total utility function. This notion implies choices bear a certain probability and uncertainty, which may distort the choice prediction.

$$U_{in} = V_i + \varepsilon_{in} = \sum_m \beta_m \cdot x_{im} + \varepsilon_{in} \tag{2.1}$$

where:
i,j = Alternative-subscripts
m = Attribute-subscripts
x = Attribute-values
β = Taste / weight
ε = Randomness
V_i = Systematic utility

Choice probabilities

The overall utility is composed of a systematic utility and randomness. The generation of choice probabilities characterises the probabilistic feature of choice models. Hence, this reflects that decision-makers are inconsistent in their behaviour, and the choice model does not perfectly represent the choice behaviour relevant for every single choice (McFadden et al., 1973). The error term prevents the least popular alternative from having a zero choice probability and the most popular to capture perfect fit. The closed-form choice probability formulation is illustrated hereafter and constitutes the probability of an alternative chosen in the linear-additive multinomial logit model

(MNL).

$$P(i) = P\left(V_i + \varepsilon_i > V_j + \varepsilon_j, \forall j \neq i\right) = \frac{\exp\left(V_i\right)}{\sum_{j=1..J} \exp\left(V_j\right)} = \frac{\exp\left(\sum_m \beta_m x_{im}\right)}{\sum_{j=1..J} \exp\left(\sum_m \beta_m x_{jm}\right)} \tag{2.2}$$

where:
P_i = probability that alternative i is chosen
V_i = systematic utility of alternative i

The choice probabilities are subject to the size of the choice task. The choice tasks are usually binary or trinary, in which the latter provides more information for each choice than the former. The model is estimated by iteratively finding combinations for all parameters βthat are most likely the data.

$$LL(\beta) = \ln\left(\prod_n \prod_i P_n(i \mid \beta)^{y_n(i)}\right) = \sum_n \sum_i y_n(i) \cdot \ln\left(P_n(i \mid \beta)\right) \tag{2.3}$$

2.3 BAIT as a DSS

Given the employment of conventional intelligent DSS, it is noteworthy to explain how BAIT is constructed. Councyl developed a new approach to decision-making tools by designing a DSS that is founded on DCM and aims to reveal how decisions are made given the explicit criteria. This section explains the approach and design phases.

2.3.1 BAIT approach to decision support

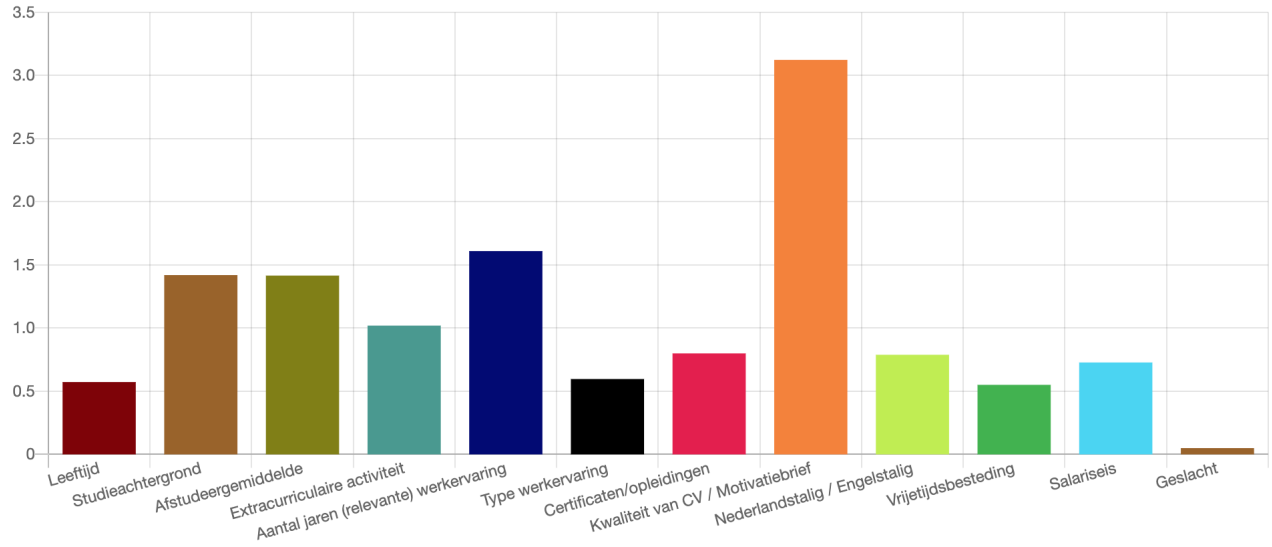
The main goal of BAIT is to present experts with their combined expertise in the context of a specific decision-making (ten Broeke, Hulscher, Heyning, Kooi, & Chorus, 2021). BAIT uses the choice modelling technique, which enables to identify preferences of large populations for commercial purposes or public policy (e.g. whether or not to construct a new highway, determine the price of public transport, etc.) (McFadden, 2001). The execution of a choice experiment is a suitable method to acquire the preferences of every individual respondent by choosing between various alternatives, each containing its combination of varying attribute-levels (Merino-Castello, 2003). The estimation of a choice model on respondents' preferences is input for calculating the relative importance of each criterion. On that account, respondents' choice information serves to calculate choice probabilities for future decisions.

The primary purpose of a choice experiment is to elicit the preferences of individuals on decision-making, as research proves that it is not very easy to explain the logic behind decisions (Hensher & Assoiate, 1993). By conducting a choice experiment and creating the model, BAIT supports the improved understanding of experts on their decision-making by obtaining their decision rules. This does not only provide insights on how decisions are made but also initiates discussion among them to improve their decisions over time (ten Broeke et al., 2021). Once the model is built, its advice (i.e. choice probability) can be improved over time by its self-learning mechanism based on future decisions. Consequently, this enables the experts to evaluate their choices over time and identify patterns in their decision-making.

2.3.2 Interface of BAIT

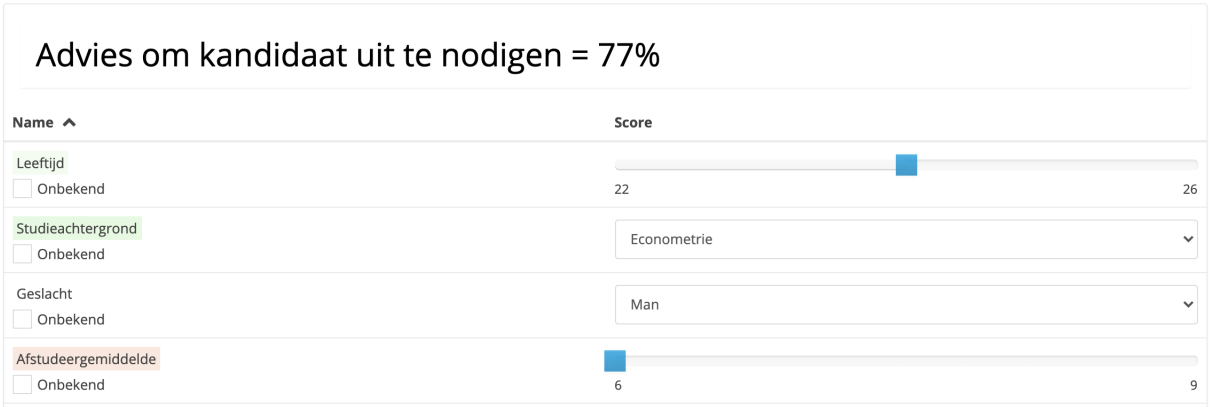
BAIT contains various screens with information about the criteria and weights. The calculated weights for all involved attributes are normalised to obtain the relative importance (ten Broeke et al., 2021). Exemplary is the relative importance data of the Human Resource demo of Councyl shown in fig. 2.4. This figure aims to provide the client explanations on the decisions and support them in the future. Due to its real-time character, the weights are determined based on the choice experiment and may vary over time by supplying the model with new information (e.g. actual choices). Moreover, the weights of the model provide the expert with insights into their behaviour. The expert may use this information to improve its way of reasoning or re-think its end-decisions. The real-time characteristic represents an internal feedback loop between the expert and the model and therefore is up-to-date at any given time.

Figure 2.4: Relative importance of HR demo



The weights play a role when future cases are provided to the model. Hence, an HR expert could fill in the information of a new applicant and retrieve probabilistic advice on whether or not to invite the applicant. The input for the real-time character is represented in this part, as the expert will fill in new input values for each given attribute and eventually decide whether or not to accept the advice of the model. The interpretation of acceptance is highly dependent on the boundary of the probability rate. A pre-determined boundary for the probability rate is crucial for the expert's interpretation (e.g. 80%). If the probability of the new input situation exceeds the pre-determined boundary, the expert may notify whether the advice is adopted. An incomplete illustration is shown in fig. 2.5.

Figure 2.5: Model input example



2.3.3 BAIT design phases

The design process requires intensive collaboration between one of the clients, as mentioned above and Councyl. The project entails three phases in order to set up a working DSS. First, the problem owner will set up the choice experiment. The choice experiment will be conducted, and the model will be estimated. Afterwards, the model will be validated and applied. We clarify the entire pilot in this section to give an overview of the work method.

Phase 1: Setting up choice experiment

The first phase entails the design of the choice experiment. This is an indispensable part in order design a choice behaviour model in a later stage: (1) construct the choices within the given

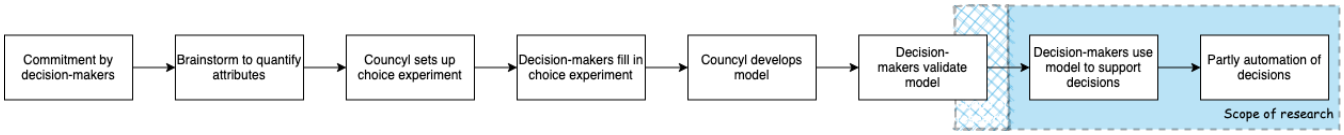


Figure 2.6: Phase scope of research

context, (2) determine crucial criteria that dictate the choice, (3) identify personal characteristics of respondents that may play a role and (4) determine the validation method.

The first step within this phase entails determining complex and straightforward cases based on a multitude of requirements. This classification step enables Councyl to design distinctive models based on any feature (e.g. complexity) if desired. The aim is to conceptualise the decision with the leading main research question. Accordingly, the choice types are determined within this step to understand the decision-making more profoundly.

After the choice types are determined, the problem owner determines the essential criteria that shape the decision. By the hypothesis of behaviour (i.e. that individuals classify their choices in terms of the stated criteria), it is possible to reduce a large set of possibilities to a smaller number (Swait & Ben-Akiva, 1987). Hence, choice modellers aim to extract the domain knowledge of experts and understand the trade-offs. Potential criteria that play a role in the decision-making are identified through interviews with various experts. The criteria have to meet several prerequisites to include in a choice experiment. The criteria should be quantified numerically, binary or categorically, and the criteria ought not to have too much overlap with each other. The latter is necessary to make clear distinctions between the criteria and potential trade-offs.

Based on the output of the stated choice experiment, it is possible to retrieve the personal characteristics of the attendees, e.g. seniority, expertise, work experience, etc. Councyl can predict various models based on personal traits and hence obtain additional knowledge. This step provides further insights on how group characteristics possibly affect the trade-offs and criteria. Accordingly, it is possible to assign different weights to criteria if significant differences are measured and if considered necessary.

Based on the data obtained from the stated choice experiment, the first validation can be executed. Within this step, choice modellers assess to what extent the model can reproduce the made decisions. The additional insights and usefulness for the decision-makers will be assessed to measure the level of intuitiveness. In addition, the accuracy and usability of the model will be evaluated in two ways: retrospectively (i.e. testing the model based on labelled historical data) and prospectively (i.e. running the model on the background on actual decisions made in the future).

Phase 2: Choice experiment and model estimation

Based on the provided criteria and range boundaries, Councyl submits the stated choice experiments to the group of experts. The choice experiment supports predicting the effects of all different criteria (or combinations) based on a minimum number of choice sets. The group of experts is selected based on their expertise to ensure qualitative data as input for the model. The stated choice experiment consists of fictive scenarios in which experts are asked to choose one of the presented alternatives within the given choice set. Such scenarios are defined in terms of scores for each criterion. Experts have to decide which alternative they prefer. The number of choice sets that each expert has to fill in varies between 20 and 40. This number depends on the number of criteria.

Based on the given answers by the experts, Councyl predicts the model as such that the criteria and weights reflect the choices. Subsequently, Counyl presents the model and additional insights to the clients with a report. According to the results, the experts can conclude whether the model correctly presents the implicit trade-offs and provides the correct output intuitively.

Phase 3: Validation and application

In the last phase, Councyl validates the model to construct a reliable DSS for potential usage. The pre-determined validation method will be applied within this phase. Based on the data, Councyl will determine the accuracy of the models using various metrics. Cooperatively, Councyl and the clients will discuss how the interactive model can provide additional value to the decision-making as a DSS.

When the clients validate and accept the models, the model can function as a DSS to special decision-making assistance. By doing so, the experts may invoke the DSS to support their decision-making. The consecutive step is to apply the DSS on a broader scale. Based on the given input, the application will provide the percentage of experts who would decide to accept the request. The application provides colour codes to prove the extent of each criterion's positive or negative effect. The expert may choose to save the decision made for each application, which has two functions for further improvement. First, automatic monitoring will be enabled by which the expert may evaluate their decision. Moreover, the expert may assess the decision of the model. This function contributes to the progress of the model itself and the improvement of choices made by experts. Second, the model contains a self-learning aspect to reflect the knowledge of the experts increasingly.

2.4 Task allocation

The level of automation of a DSS highly affects the available space for freedom of choice for the end-user. Parasuraman, Sheridan, and Wickens (2000) proposes a scale on Level of Automation (LOA) of human-machine systems that can be applied on four broad classes of functions: (1) information acquisition; (2) information analysis; (3) decision and action selection; and (4) action implementation. The computer function and human role are illustrated coherently in fig. 2.7. The higher the level of automation is, the less potential space is left for the human to control the decision. Hence, the figure does not imply a negative correlation between the levels of automation and human autonomy. By contrast, it co-determines the available room for HMA given the LOA. It is used for indicative purposes to illustrate the interplay between the automation of digital systems and the user's autonomy.

Figure 2.7: Level of automation

(Parasuraman et al., 2000)

Autonomy level human agent	Automation specification and artificial agent role
High human moral autonomy	1. The computer offers no assistance; the human must take all decisions and actions. 2. The computer offers a complete set of decision/action alternatives. 3. The computer narrows the selection down to a few.
Partial agent autonomy	4. The computer suggests one alternative. 5. The computer executes that suggestion if the human approves. 6. The computer allows the human a restricted time to veto before automatic execution. 7. The computer executes automatically, then necessarily informs the human.
Low human moral autonomy	8. The computer informs the human only if asked . 9. The computer informs the human only if it, the computer, decides to. 10. The computer decides everything and acts autonomously, ignoring the human.

BAIT

The categorisation of BAIT on the LOA scale is essential for multiple reasons. Most importantly, O'Neill, McNeese, Barron, and Schelble (2020) states the interaction between the artificial agent and human agent can be considered as a Human-Autonomy Teams (HAT) if the artificial agent exceeds level 4 on the LOA scale. This level implies coordinated and meaningful cooperation on a joint task between the two entities. Lower levels of automation merely indicate simple automation of a task by the machine but cannot be considered an actual entity that is viewed as a team member. Two overarching criteria are identified for a device to be considered a team member (i.e. an autonomous agent). First, a degree of interdependence with other team member activities and outcomes is required for recognition (Walliser, de Visser, Wiese, & Shaw, 2019). Second, a degree of agency involving independent actions by the artificial agent is an essential characteristic as this is a crucial proxy that reflects the degree of automation (Wynne & Lyons, 2018). Altogether, the artificial agent can be classified as being "autonomous" when recognised as a distinct team member in the execution of a decision-making process (O'Neill et al., 2020).

The system can conduct a trade-off between predefined criteria, which are defined by weights. Hence, by definition, BAIT can act minimally on level four given its system characteristics and the type of information it displays. This capability implies the system can suggest one alternative. However, the system is potentially able to act on a higher LOA. If integrated properly, the system can execute the suggestion with or without approval by the end-user. This range of automation levels implies the end-user may at least get rid of the conventional think process in the decision-making. The end-user is only expected to assess the suggestion and understand the input values shallowly. However, if BAIT acts on the seventh level of automation, the system will only inform the human of the decision. In this case, the end-user still can intervene in the decision-making but has less time to do so. To conclude, the task allocation between the human and BAIT varies extensively. BAIT can act on different automation levels, by which the human has more or less hypothesised space of autonomy. Depending on the decision-making, case complexity and context, end-users may utilise BAIT on varying automation levels.

Summary box chapter 2

The goal of chapter 2 is to define BAIT as a technology. On that account, we were able to put it in the spectrum of decision-making technologies and explain its fundamental characteristics. A profound literature study and analytical comparison with other technologies enabled us to define the technology more precisely. The following points summarise this chapter:

- BAIT is considered an intelligent decision support system (IDSS) that acts similarly to expert systems. BAIT does this by transferring 'human' expertise into digital systems to treat repetitive decisions.
- BAIT is by definition a knowledge-based system with non-knowledge based peculiarities. It uses statistical computations to find patterns based on specific, case-dependent choice modelling data.
- BAIT uses the discrete choice modelling (DCM) technique to quantify choices into specific decision criteria.
- The construction of the DSS happens in three distinctive phases: setting up the choice experiment, conduct the choice experiment & model estimation and validate & apply the model.
- BAIT provides advice to the end-user by presenting the relative importance between criteria. Moreover, it shows the end-user the percentage of peers that would likely act upon a specific case. This case consists of certain attribute-level combinations.
- We pinpoint BAIT on the level of automation (LOA) scale between levels four and seven. This range implies a partial agent autonomy, in which the end-user and the technology perform distinctive tasks.

Chapter 3

Theoretical framework: operationalising human moral autonomy

> "Technological artifacts are not neutral intermediaries but actively co-shape people's being in the world: their perceptions and actions, experience and existence... When technologies co-shape human actions, they give material answers to the ethical question of how to act."
>
> Verbeek, PP, Moralizing Technology: Understanding and Designing the Morality of Things

In this chapter, we study the definition of moral autonomy, what it includes and how it manifests within the context of BAIT. In section 3.1 we argue the importance of HMA, provide a characterisation of this philosophical concept and explain the effects of automation. After that, we define the theoretical constructs of HMA in section 3.2. Those steps enables to translate the theoretical constructs into measurable variables in a survey. By doing so, we aim to formulate an answer to the following question:

Research question 2

What is a philosophical definition of human moral autonomy in the context of a DSS like BAIT?

3.1 Characterisation of Human Moral Autonomy

Within this section, the objective is to characterise HMA in the context of BAIT. In section 3.1.1 the importance of moral autonomy is argued by substantially proving the significant role this ethical value fulfils. Section 3.1.2 aims to give - based on a multitude of theoretical foundations - a suitable definition of moral autonomy for this specific research.

3.1.1 Importance of Moral Autonomy

Providing sufficient information on system characteristics and the decision-making is paramount to prevent erosion of HMA (AI HLEG, 2019). Such decays can be controlled by governance mechanisms such as the human-in-command (HIC) approach. This section will elaborate on the importance of moral autonomy and provide arguments for enhancing human autonomy in the employment of BAIT.

Information systems for decision-making were initially build to improve human reasoning employing digital computational devices (Van den Hoven, 1998). The concept of improving human intellectual functioning utilizing computers is defined as epistemic empowerment. The replacement of traditional methodologies by artificial systems affects our ideal of intellectual autonomy and individual responsibility (Nissenbaum, 1996). Hence there is, as Arrow (1974) argues, a trade-off between relying on an artificial authority and responsibility of the individual. Additionally, the depletion of the decision-maker's moral agency could result in the ability to harm people using psychological distancing (Cummings, 2006). The question then arises to what extent decision-makers can be held responsible for their actions, given their reasoning based on the limited information provided by digital systems.

The moral autonomy of humans is desirable for many reasons. Firstly, present-day machines are not capable of making moral distinctions (Santoni de Sio & van den Hoven, 2018). The urgency arises to acknowledge and invalidate the assumption that devices contain sufficient moral sensitivity to automate decisions completely. Moreover, the lack of moral autonomy embodies a significant role in poor design decisions of DSS, which may cause issues on responsibility and accountability (Cummings, 2006). A various number of issues might appear accordingly. One of them is the incorrect allocation of moral agency from humans to computers. Additionally, highly automated DSS may cause humans to view the system as an independent agent that is capable of wilful action (N. Sarter & Woods, 1994). An increased focus on the moral autonomy of humans enables the prevention of such states wherein humans do not reckon themselves to be responsible and accountable for their actions.

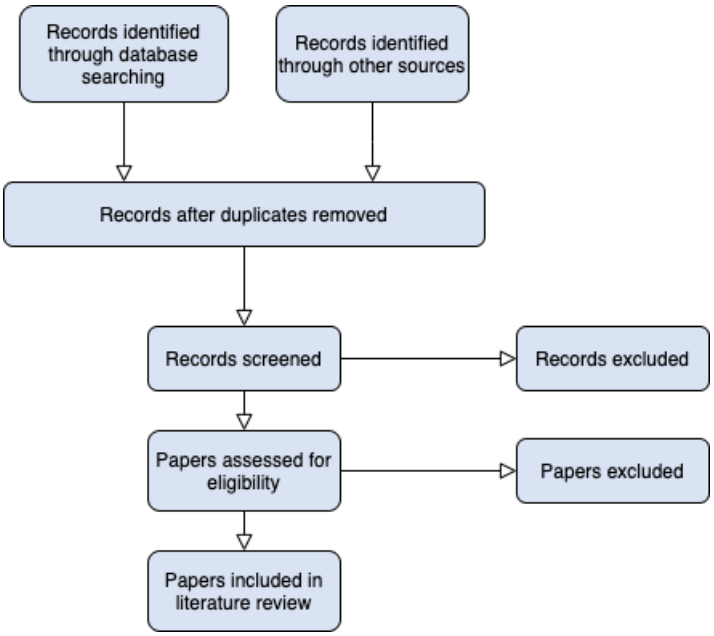
Ethics-based auditing of moral autonomy in BAIT may potentially support the authority of the end-user. (Mökander & Floridi, 2021). Firstly, it may support the DSS by visualising and monitoring outcomes more extensively. Increased explainability also enables the decision-makers to formulate sufficient arguments and prove their reasoning. One can increase explainability by providing the decision-maker with additional information on the algorithmic process. Furthermore, it supports explaining to end-users how a decision is made based on what grounds. This additional information will likely improve their understanding of the decision-making and legitimise the decisions more effectively. Additionally, experts still want to make their own choices; this can only be enabled when sufficient and sound information is provided. Lastly, it will help explore accountability and co-determine the allocation of responsibility by integrating the system into existing structures.

Like all solution-driven research studies, there are drawbacks to the approach and purpose of this study. On meta-level, one should be aware that the purpose of this study is not to contribute to the tendency in science and technology of "ethics bluewashing". Floridi (2019) defines ethics bluewashing as: "the malpractice of making unsubstantiated or misleading claims about ethical values of benefits of digital processes, products, services or other solutions to appear more digitally ethical than one is". On that account, we note that ethics bluewashing cannot be prevented in the long-term with certainty but must be taken into account in the execution of this study. Furthermore, committing the naturalistic fallacy can be only made as a conclusion from a discussion and should not be used as an instrument for deciding it (Frankena, 1939). The is-ought fallacy illuminates the distinction respectively between facts and values, between the descriptive and normative. To conclude, this fallacy is highlighted by philosophers in numerous studies. They all indicate avoiding such practices at all costs.

3.1.2 Defining Human Moral Autonomy

The purpose of this section is to define a suitable concept of HMA to assess HMA both empirically and ethically. Moral autonomy is defined in diverging ways by numerous philosophers, implying they do not represent identical meanings (Dworkin, 2015). Moral autonomy is within this study conceptually confined to the autonomy of humans, as opposed to machine autonomy, which is the opposite of the same dichotomy. On that account, the plurality of definitions of moral autonomy demands a concretization. Therefore, influential papers on moral autonomy and digital environments are collected to form an appropriate definition for this study, as illustrated in fig. 3.1. A number of noteworthy definitions of moral autonomy, collected by Dworkin (2015), are presented in table 3.1.

Figure 3.1: Literature review process



The technological advancement of automated decision-making tools induced a two-folded linguistic issue: technologies are increasingly defined in terms of autonomy while humans are gradually more defined in terms of automation (Chiodo, 2021). Unsurprisingly, the two concepts of autonomy and automation insinuate two diverging meanings. Kant (1785, p.44) defines autonomy as "self-given law" and constitutes "autonomy of the will to be the property of the will by which it is a law to itself". In contrast, current technological automation implies a notion of automation that is characterised as "self-given" that is something "off-hand" (Chiodo, 2021). While autonomy and automation both do not represent the exact definition, they are used interchangeably for complete separate entities like machines and humans. Woods (1996) suggests the automation development induces an imprecise substitute in the perception of automation of systems towards autonomous machine agents instead. Hence, this study's definition of moral autonomy is explicitly entitled as human moral autonomy (HMA) since merely moral autonomy of humans, i.e. decision-makers, will be considered.

In contrast to the notion of Chiodo (2021), other researchers state artificial agents can be perceived as moral agents given the context in which the observable entity is evaluated. Floridi and Sanders (2004) defined three criteria for a level of abstraction to classify an entity as a moral agent: (1) interactivity (agent and its environment can act upon each other), (2) autonomy (agent can change state without direct response to interaction) and (3) adaptability (agent's interactions can change the transition rules by which it changes states). The formulated criteria are to some degree interdependent and entirely embody a moral agent. Given these criteria and the context in which BAIT will operate, the artificial agent (i.e. DSS) can be held accountable for moral actions according to this line of reasoning. Nevertheless, this perspective does not affect our conceptualization of HMA.

The definition of HMA is context-dependent, which emphasises the urgency to define the context. Human-computer interaction implies a distributed allocation of moral supervision and work supervision between the human agent and the artificial agent (Waa & Diggelen, 2020). The allocation distribution is highly dependent on the degree of automation. Moral supervision entails recognising the situation, identifying moral dimension and deciding on significance (Sternberg, 2012). Hence, moral supervision implies the human agent is not doing all the work, but strictly meaning observing the decisions by an artificial agent. Whenever it is deemed necessary, the human agent could take over the decision-making. The subdivision of moral supervision is illustrated in fig. 3.2.

Dworkin (1981) formulated six specific characterisations that underline the generic formulation of moral autonomy. The characterisations embody the meaning of possessing ethical principles

Table 3.1: Overview of definitions on (moral) autonomy

Definition	**Source**
The law in thus implementing its basic commitment to man's autonomy, his freedom to and his freedom from, acknowledge(s) how complex man is.	(Goldstein, 1978, p. 252)
To regard himself as autonomous in the sense I have in mind, a person must see himself as sovereign in deciding what to believe and in weighing competing reasons for action.	(Scanlon, 1972, p. 215)
As Kant argued, moral autonomy is a combination of freedom and responsibility; it is a submission to laws that one has made for oneself. The autonomous man, insofar as he is autonomous, is not subject to the will of another.	(Wolff, 1998, p. 14)
(Children) finally pass to the level of autonomy when they appreciate that rules are alterable, that they can be criticized and should be accepted or rejected on a basis of reciprocity and fairness. The emergence of rational reflection about rules ... central to the Kantian conception of autonomy, is the main feature of the final level of moral development.	(Peters, 1972, p. 130)
I am autonomous if I rule me, and no one else rules.	(Feinberg, 1982, p. 161)
Human beings are commonly spoken of as autonomous creatures. We have suggested that their autonomy consists in their ability to choose whether to think in a certain way insofar as thinking is acting; in their freedom from obligation within certain spheres of life; and in their moral individuality.	(Downie & Telfer, 1971, p. 301)
A person is "autonomous" to the degree that what he thinks and does cannot be explained without reference to his own activity of mind.	(Dearden, 1972, p. 453)
Acting autonomously is acting from principles that we would consent to as free and equal rational beings.	(Rawls, 1971, p. 516)
I, and I alone, am ultimately responsible for the decisions I make, and am in that sense autonomous.	(Lucas, 1966)

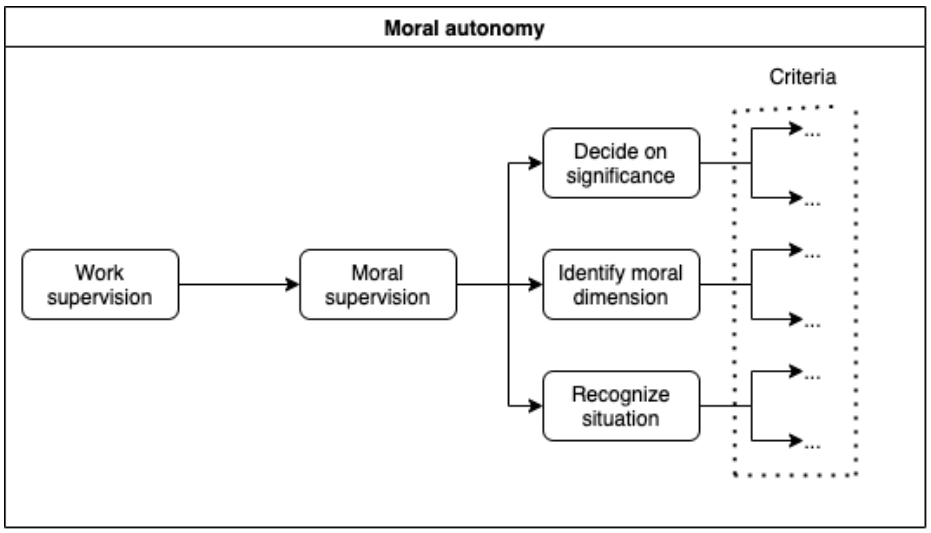


Figure 3.2: A model for ethical reasoning (Sternberg, 2012)

and acting upon them accordingly. We decided to use one definition to build further on the theory of moral autonomy: "A person is morally autonomous if and only if he bears the responsibility for the moral theory he accepts and the principles he applies" (Dworkin, 1981, p. 30). This characterisation contains two crucial aspects which make it suitable. First, it entails bearing responsibility for the moral theory the human agent accepts for the decision that has been made. Hence, the human decision-maker must understand the AI's moral theory or decision principle to support decisions. Second, characterisation suggests the human agent applies moral principles in making decisions. We can assess those principles to understand how much the human complies with their principles in making choices.

We can define those principles by extending the pre-defined characterisation of moral autonomy by Dworkin (1981). This is filled in by the principles of moral supervision defined by Sternberg (2012) and given in fig. 3.2. These principles for moral supervision are defined within the context of the interaction between an artificial agent and a human agent (Waa & Diggelen, 2020). The acknowledgement of a dynamic allocation of tasks for moral decision-making in a human-agent context is highly valuable within this research. To conclude, this state-of-the-art concept enables the researcher to actively study the degree of HMA in the context of HCI.

The objective is to formulate a definition of HMA based on a broad set of definitions. By doing so, we aim to formulate a comprehensive definition within the scope and purpose of this study. Therefore, the characterisation as mentioned above of moral autonomy defined by Dworkin (1981) and the principles of moral supervision defined by Sternberg (2012) will be combined to set the boundaries of this research given the definition. Whereas Dworkin (1981) merely mentions "responsibility" and "principles" to define moral autonomy, Sternberg (2012) emphasises on the authority of the end-user and the execution of moral supervision on the system. Moreover, Sternberg (2012) distinguishes moral supervision in terms of the situation (i.e. having sensitivity for the context) and the DSS itself. The synthesis leads to the following definition: "*A person is morally autonomous if and only if he bears the responsibility and authority for moral supervision on the situation and the decision support system*". The definition can be decomposed as follows. A person is considered morally autonomous if he bears the responsibility for the decisions made using the DSS. Moreover, he is expected to be in charge of the decision-making; this is what we refer to as authority. The mentioned moral supervision is previously explained and illustrated in fig. 3.2: it entails the decision on significance, identification of ethical dimensions and recognises the situation. Lastly, the definition ends with the acceptance of the decision-maker with the moral principles of moral supervision.

3.2 HMA framework: operationalising the concept

In section 3.1 we defined the concept of HMA in the context of BAIT. Accordingly, we can operationalise the concept into theoretical constructs. Hence, this section clarifies the operationalisation and interpretation of our understanding of HMA. In section 3.2.1 it is explained how the theoretical framework is constructed, in section 3.2.2 the theoretical constructs are independently defined and in section 3.2.3 the relationships between the constructs are explained through a causal diagram.

3.2.1 Construction of framework

We use the literature regarding HMA to construct a theoretical framework. Although the available literature provides sufficient insights on the concept of HMA, it needs to be translated to constructs to assess it empirically. Constructs are the foundations of the scientific theory and enable measurable variables (van der Waa, Nieuwburg, Cremers, & Neerincx, 2021). We do this by using well-defined constructs, draw a causal diagram to show the relation between them, and building upon existing theories.

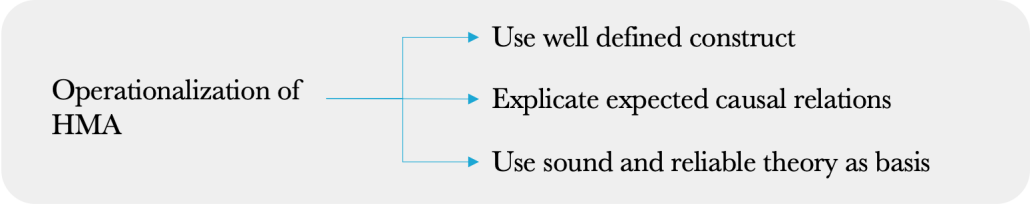


Figure 3.3: Operationalisation of HMA theory
Derived from (van der Waa et al., 2021)

3.2.2 Defining the constructs

This sub-section covers the characterisation of the constructs. The constructs within this study are of two intermediate types: requirements (defined by Waa and Diggelen (2020)) and conditions (defined by Van den Hoven (1998)). Consequently, the constructs can be designated to specific questions in the questionnaire for measurement purposes.

Team design patterns

The article, written by Waa and Diggelen (2020), prescribes different requirements for various Team Design Patterns (TDP). We decide to consider TDP2 as an ideal setting of BAIT where decision-makers will co-operate with the DSS. This particular setting is called "Supported moral decision making". It is characterised by an artificial agent (i.e. the DSS) performing the main task autonomously, while the human agent solely supervises the moral sensitivity of a situation. TDPs 2, 3 and 4 perfectly fit into the LOA scale of level 4, and higher presented by Parasuraman et al. (2000) and illustrated in fig. 2.7. They represent Human-Autonomy Teams (HAT) in various distributions of task-allocation between the human and artificial agent. Allocation of moral competencies within TDP2 shows overlap to a great extent with a setting in which BAIT will operate as a DSS. When the need is identified for a moral decision, the human agent may take over. Therefore, the DSS supports the human agent by providing explanations and other relevant points to make a moral decision. The following requirements are set for this particular setting between the artificial and human agent, which we illustrate in fig. 3.4:

1. **Time:** the human agent must predict morally sensitive decision in time. While the artificial agent performs the main task autonomously, the human agent supervises the findings and must assess moral sensitivity in each case with a particular time constraint.
2. **Moral implications:** the human agent is expected to have a sufficient understanding of the moral implications given for each case that has been treated by the artificial agent. This notion requires situational awareness and moral sensitivity.
3. **Moral context:** the artificial agent must explain the moral context sufficiently to enable the possibility as mentioned earlier of making moral decisions by the human agent. Hence, the human agent can only be expected to make an ethical decision if the artificial agent provides sufficient context.

4. **Halt/resume:** the system must include the capability to cease or continue the decision-making process at any given time if this is deemed necessary.

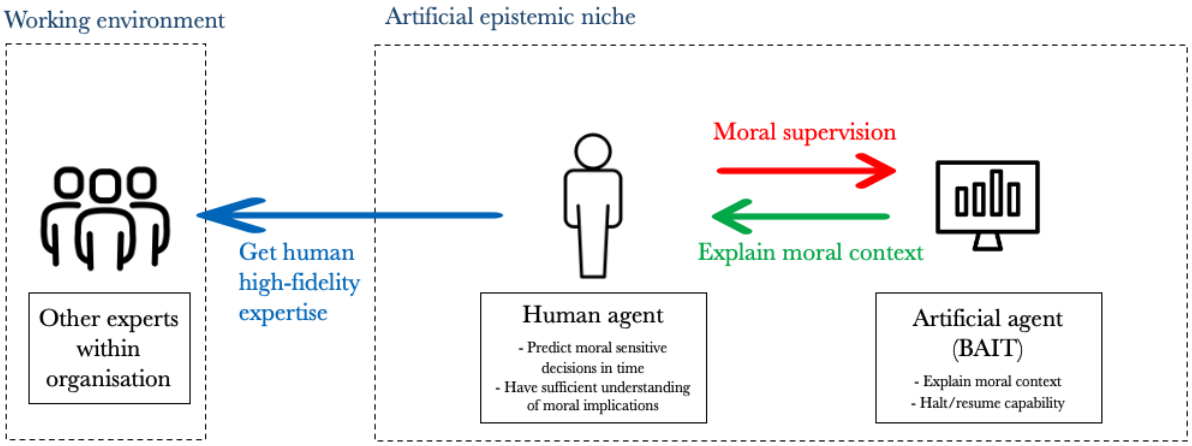


Figure 3.4: Task allocation of BAIT
(Waa & Diggelen, 2020)

Conditions for HMA

The second source is the paper of Van den Hoven (1998). Computerised work environments are knowledgeable entities within professional groups and function as centres of coordinated action by users (Van den Hoven, 1998). We refer this description to as 'artificial epistemic niches'. These artificial epistemic niches may cause narrowly embedded systems. Narrowly embedded systems are characterised by a lower ability for the human agent to critically and morally reflect on the system. When the situations meet the conditions, Van den Hoven (1998) argues humans lose their status as autonomous moral persons. On that account, the conditions determine the extent to which a human agent is stuck in the narrowly embedded system. More importantly, the literature considers the conditions to be a proxy for HMA in a computerised environment. To conclude, if the conditions are met, we believe the human agent is stuck in the narrowly embedded system. As a consequence, the human agent has a low autonomy in decision-making.

1. **Inscrutinizability:** exploring the internal machine operation is an essential feature for multiple reasons. Firstly, the accessibility of systems enables to monitor the decision-making process performed by the DSS. Second, the tractability of operations within a system provides the ability for human agents to trace back decisions and independently draw conclusions.

2. **Pressure condition:** the execution of complex real-time decisions reliably. This condition entails three sub-conditions: (1) the times-span for decision-making is limited, (2) decisions of a specific type or domain have to be taken and (3) the inability to get additional support from fellow human agents. If all three sub-conditions are met, one can state the pressure condition is in effect.

3. **Error condition:** stupidities, inconsistencies, appearance of bugs and corrupted data characterise the error condition. Literature divides this situation into five sub-conditions under which this condition may occur: (1) flaws in the specification, (2) brittleness of digital systems, (3) bugs and programming errors, (4) limits of testing and proof and (5) emergent and unpredictable properties of software may occur in the integration of multiple systems.

4. **Critical questioning and evaluation:** given the first three conditions, they collectively weaken the critical reasoning of human agents as a result of lacking expert opinions on decision-making processes.

3.2.3 Causal diagram

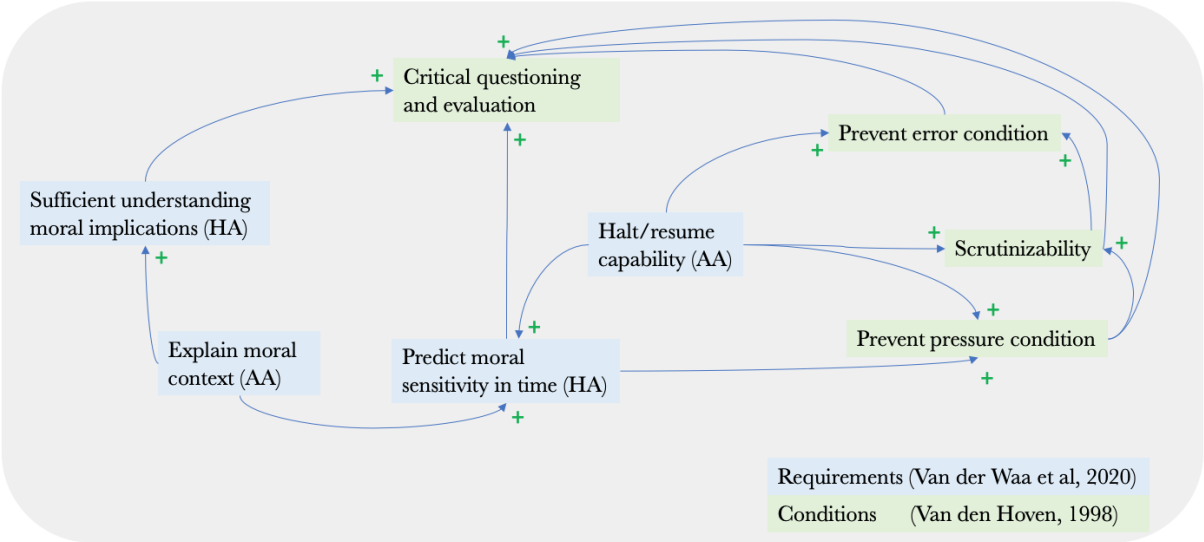
This subsection contains and explains a causal diagram to present the expected causal relations between constructs (Pearl et al., 2009). We use the connections to formulate hypotheses in terms

of the constructs, preserve theoretical consistency and provide additional insights on the results related to the constructs (van der Waa et al., 2021). This methodology is in line with the second sub-recommendation of the first recommendation illustrated in fig. 8.1.

The causal diagram illustrated in fig. 3.5 contains the requirements defined by Waa and Diggelen (2020) and the conditions defined Van den Hoven (1998) and the relations between the constructs will be explicated hereafter. The requirements, to start with, are separable in task allocation between the human agent (HA) and artificial agent (AA). Sufficient understanding of moral implications is an important requirement for the HA within the computerized environment and mostly affects the capability to question and evaluate decisions critically. Moreover, the HA must be capable of predicting the moral sensitivity of a case in time. This requirement will potentially affect the capability to question and evaluate decisions critically but is also causally related to preventing the pressure condition. We assumed this based on the pressure condition's sub-conditions that contain the time-span of a decision, the type of decisions, and the ability to get support externally. Hence, moral sensitivity is assumed to be predicted correctly when the pressure condition is non-existent in a situation. The AA also bears task responsibility that affects moral decision-making. First, the AA must explain the moral context sufficiently to enable the HA to understand the moral implications and predict the moral sensitivity in time. The halt/resume capability of the AA affects many other constructs. It prevents the error condition for evident reasons, enables the HA to scrutinize through the DSS, reduces the pressure condition, and enables the HA to predict moral sensitivity in time.

The second paper integrated into the causal diagram are the conditions defined by Van den Hoven (1998) and are treated as metrics for the empirical study. The first condition, preventing the pressure condition, affects scrutinizability with the assumption that the HA is only able to scrutinize through the system if the HA is not exposed to one of the sub-conditions representing the pressure condition. Scrutinizability has a determining influence on preventing the error condition, as one can only find flaws, bugs or any other inconsistency when he can inspect the system. Lastly, as Van den Hoven (1998) states in his paper, given the first three conditions, information systems are sub-optimal for the capability of critical questioning and evaluation by the HA. Hence, if all conditions are prevented, critical and moral reasoning would lead to a greater extent.

Figure 3.5: Causal diagram

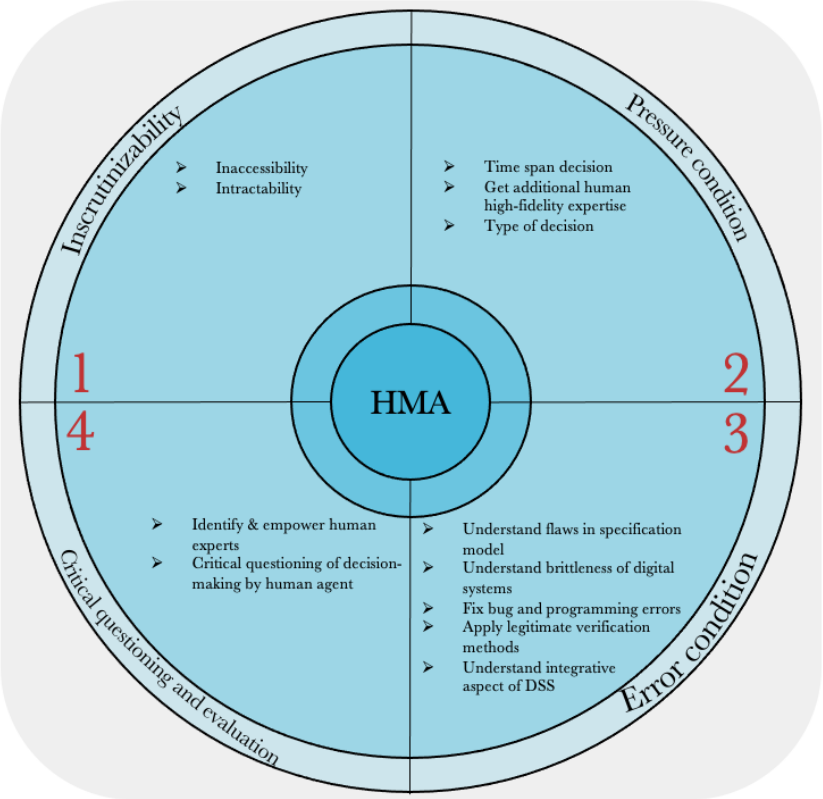


3.3 Outcome: the HMA Framework

After we define the constructs and the causal relations between them, we are able to build accurately on existing theories. Waa and Diggelen (2020) recommends adopting existing theory from various fields such as philosophy and human-computer ineteraction (HCI) to provide explanations.

As mentioned before in section 3.2.3, the conditions of Van den Hoven (1998) will be used accordingly within this chapter to set up the constructs for the questionnaires. Constructs represent separable components of a condition and will be used to formulate the questions in the questionnaire. Therefore, the conditions are translated into constructs to set up measurable items. The conditions and items are shown in fig. 3.6.

Figure 3.6: Theoretical framework



The scrutinizability condition characterises the first part of the framework. This condition can be subdivided into the accessibility and tractability of a DSS. Accessibility entails the ability to monitor the internal process within a system. Therefore, it implies the extent to which the HA can access additional information on the system itself. Tractability is - under the pre-condition of having access to the system - keeping track of the decisions. Hence, the combination of accessibility and tractability is foundational for determining the degree of scrutinizability of a system. To conclude, within this part, the framework will cover statements regarding both concepts of accessibility and tractability to assess the degree of scrutinizability.

The second part of the framework covers the pressure condition. This condition entails the time to decide, getting human advice and the type of decision. The first sub-condition aims to measure the period for each decision and assess whether the time is sufficient, given the kind of decision. The second sub-condition evaluates the ability to get support from a human colleague, preferably outside the concerned DSS. The last sub-condition assesses the influence of the types of decisions, i.e. complex or straightforward case, to measure the degree of feeling a pressure given the first two sub-conditions.

The third part of the framework encompasses the error condition. This condition is sub-divided into five sub-conditions which will be explained accordingly. The first sub-condition aims to assess the extent to which the HA understands potential flaws in the model. The second sub-condition evaluates the degree to which the HA acknowledges the brittleness of a system. This definition implies the HA is aware of the consequences due to changes in the system. The third one refrains fixing bugs and programming errors. Agreeably, we will not expect the HA to fix bugs and programming errors but identify weak spots. The fourth sub-condition entails the application of legitimate verification methods and the acknowledgement of testing and proof. Lastly, we expect

the HA to understand the role of a DSS and how it contributes to their decision-making.

The last condition is the overarching one for all conditions mentioned above: critical questioning and evaluation of systems. This condition entails the identification and empowerment of human experts as a first sub-condition. Second, it assesses (subjectively) the ability of a HA to question the decision-making critically. This overarching condition will, therefore, measure the extent to which the HA loses (or not) its position as an expert by assessing the competencies that are important for critical reasoning.

Summary box chapter 3

The goal of chapter 3 is to define and operationalise human moral autonomy (HMA) in the context of computerised environments (i.e. BAIT). We conducted philosophical research on moral autonomy and identified definitions and conditions that embody HMA more specifically. On that account, we were able to translate those conditions and prerequisites into the so-called 'HMA Framework'. This framework is the operationalisation of a comprehensive normative concept. The framework enables us to convert the theory into measurable variables in the next chapter. The following points summarise this chapter:

- The importance of HMA is multi-faceted: the literature shows a lack of moral responsibility among humans who use digital systems. Moreover, many situations prove decision-makers disproportionally view the system to be an independent agent. Lastly, most experts still want to make their own choices.
- There is a multitude of definitions of moral autonomy given by prominent philosophers.
- The definition of HMA within this study is: "A person is morally autonomous if and only if he bears the responsibility and authority for moral supervision on the situation and the decision support system, by which he accepts the principles of moral supervision which he applies."
- The definition of theoretical constructs are built on two distinctive concepts of moral decision-making.
- The HMA framework consists of one overarching concept, HMA, and four conditions. Those four conditions can be sub-divided into twelve sub-conditions. We will make use of the sub-conditions to construct the HMA survey at a later stage (see chapter 5).

Chapter 4

Research method: measuring perceptions of HMA

> "No man is morally responsible for actions unless they are performed for the sake of principles which he cannot in conscience disavow."
>
> Aiken, Reason and Conduct

This chapter clarifies the research methodology to measure the perception of HMA of experts. By doing so, we can evaluate the perceived extent of moral autonomy of end-users. We primarily do this by setting up a validated measurement instrument, the HMA survey. The survey primarily serves to answer the third sub-question. Additionally, the design of the measurement instrument enables to answer sub-research questions four and five in chapter 6. In this chapter, we present the research method of the empirical part. In section 4.1 we explain the approach for the methodology on a higher level. In section 4.2 the data collection method is explained. This section also clarifies how the data may serve various analysis purposes. After that, in section 4.3 the interview approach is explained, which we use for the pilot survey. section 4.4 explains the validation methodology. Lastly, section 4.5 shows the way the data can be used for descriptive analysis.

4.1 Research outline

The methodology of this research entails both qualitative as quantitative data collection and analysis methods within one study to understand complex phenomena (Creswell, 1999). As BAIT is still an emerging technology within the field of AI, there is little to no literature available on the user experience of this particular DSS. Moreover, there is no user data available to evaluate it accordingly. We design a survey to measure the perceptions of experts and perform a thorough analysis. The latter entails the analysis of relations between the hypothesised constructs. Surveys are suitable to obtain specific data on the constructs, in which one should consider validity, reliability and ambiguity in the data collection process (Richards & Schmidt, 2002). The validation of the survey entails two distinctive activities. Firstly, after constructing the initial version of the survey, we pre-test it by conducting interviews with potential respondents. Additionally, we use the transcription of the interviews in chapter 6 to identify patterns in the data. The second validation step is a principal component analysis (PCA) to convert the hypothesised factors. Both activities form crucial steps to validate the survey.

We use the format defined by Creswell (1999) to perform the mixed-method study. Although the entire format consists of more steps than the list below, they have been treated implicitly or explicitly in former chapters.

1. Relationship of methods to paradigms
2. Visual model of Mixed-Method approach
3. Type of research design within each method

4. Data collection within the type
    (a) **Sampling strategy and researcher's role**
    (b) **Types of data to be collected**
    (c) **Sequencing of data collection**
    (d) **Relative emphasis on qualitative/quantitative**
5. Data analysis within the type
6. Approach to validity and verification
7. Overall organization of the integrated findings and anticipated results

The visualisation of the mixed-method approach within this study is of sequential order and is illustrated in fig. 4.1. First, we design the initial survey as a measurement instrument. Subsequently, we conduct interviews as part of the pilot survey. These interviews fulfil the function of a pretest of the questionnaire. Hence, we ask interviewees questions about the theoretical constructs. After documenting the interviewees' answers, we conduct the first analysis to identify exciting output quickly. After that, we use the most interesting points from the interviews to fine-tune the questions for the quantitative part. By doing so, we aim to understand certain phenomena from the quantitative part more profoundly. Eventually, the synthesis between the two activities lead to the final version of the HMA survey. This survey will be deployed among the respondents and validated by means of factor analysis.

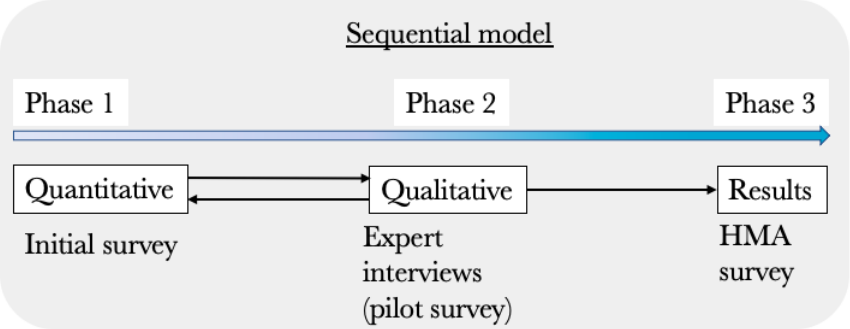


Figure 4.1: Mixed method approach

4.2 Data collection

This section explains the data collection enterprise and the sample methodology.

4.2.1 Sampling strategy and researcher's role

The sampling strategy is a crucial step in both quantitative and qualitative research methods. Trost (1986) provides a technique of seven steps to creating a varied sample for qualitative data analysis. As a general rule, such techniques could be used but will be excluded from this study for several reasons. First, this study aims to explore the perceptions of HMA in BAIT, which is a novel technology. Hence, it is hard to gain reliable information on experts' perceptions, as only a small group of people has joined the pilot with Councyl on BAIT. Second, such techniques require defining variables for the sample. This pre-selection would endanger the feasibility of this study, as it will likely exclude individuals from the small sample causing a tiny pool of potential respondents. To conclude, we encourage all individuals who participated in the choice experiments before to take part in our study.

The sample consists of the following experts that will participate in this study: surgeons and neonatologists (UMCG), consultants, public tenders (Deloitte) and intensivists (OLVG/AUMC). On that account, they will provide insights into their perceptions of HMA within BAIT given the decision-making aimed to automate. Although HMA is generically measurable throughout all domains, the possibility exists to measure different perceptions between healthcare and consultancy. Nevertheless, this is not the focus of this study, as we primarily aim to explore generic perceptions on HMA of experts within BAIT.

4.2.2 Types of data to be collected

The survey consists of statements on the conditions of narrowly embedded systems. The statements are all defined on the Likert scale, which means only ordinal data type is obtained from the questionnaire (Wu & Leung, 2017). We use the data merely for descriptive purposes, as the sample does not meet the size required to perform inferential statistical analysis. Inferential statistics demands statistical significance to make predictions or generalizations about the population. The disaggregation of descriptive statistics is illustrated in fig. 4.2 and shows the type of measures that describe the output.

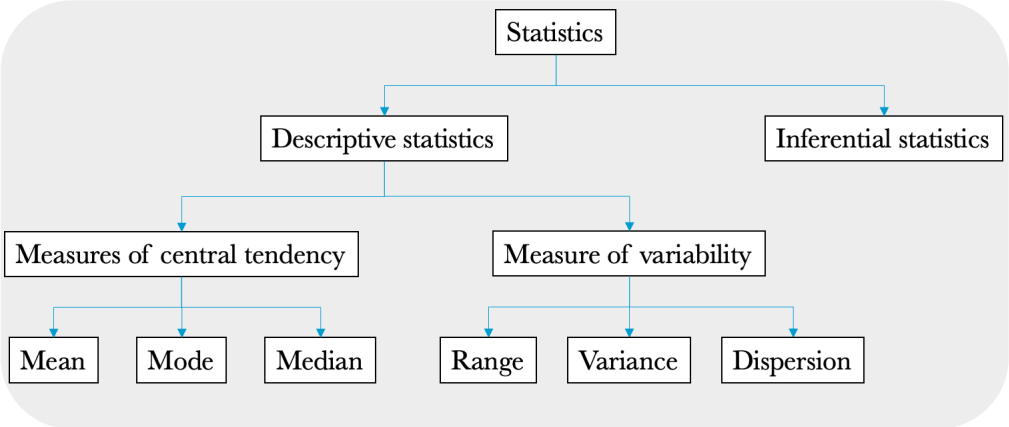


Figure 4.2: Statistics type and measures

4.2.3 Sequencing of data collection

The sequencing of the data collection begins with the qualitative part through a pilot survey, followed by the quantitative part in the form of a questionnaire. An initial survey will identify clusters of similar cases, followed by in-depth case studies of issues that represent the different cases. By doing so, we aim to get a better grasp out of the questionnaire employing the interviews.

4.2.4 Relative emphasis on quantitative/qualitative

The relative emphasis lies more on the quantitative part than on the qualitative aspect, as it reflects the overall perception of all experts on the HMA of BAIT. Hence, the qualitative part fulfils mainly a supportive function to obtain an improved understanding of the survey.

4.3 Interview approach

We mainly use the interviews to validate the questionnaire before deploying. We do this by pretesting, in which the objective is to ensure the statements within the questionnaire reflect the information as they are formulated (Grimm, 2010). Generally, two distinct lines of questions are identified with different purposes. Content mapping questions have the purpose of identifying issues to open up research, whereas content mining questions aim to explore the detail within certain subject domains (Ritchie & Lewis, 2003, p.148). The objective of this interview (i.e. pretesting) fits the definition of the latter concept. The theoretical constructs are defined beforehand, and hence the goal is to explore whether interviewees understand the constructs. A structured map is created which includes the theoretical constructs as determined in chapter 3, aligned with the recommendations of Grimm (2010). This results in the map shown in fig. 2 in Appendix A - Pilot survey.

The interview on the questionnaire consists of explanatory and clarificatory questions. We ask specific questions on constructs, starting with explanatory ones (e.g. "What do you think of your moral responsibility when using BAIT?"). After that, we asked interviewees to clarify (e.g. "In what way do you think it affects your moral responsibility?"). Explanations and examples given by Ritchie and Lewis (2003, p.151) were used to determine the structure. According to

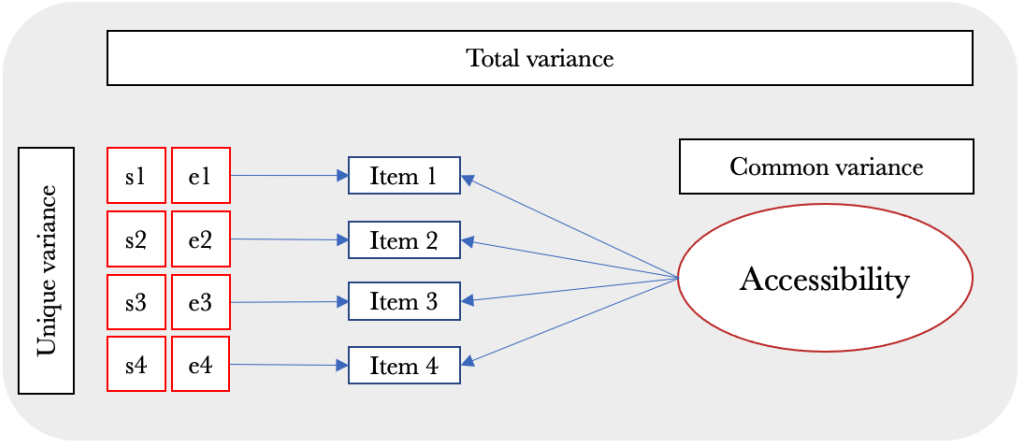
this description, the form of interchange of information that takes place is also defined as semi-structured interviews. A semi-structured interview is an interchange of verbal information in which pre-determined questions are used as the structure for an informal conversation (Longhurst, 2003). This method is deemed suitable given the limited amount of time interviewees have. Moreover, it enables us to understand potential obstacles in the survey. The documentation of the interviews with Deloitte and OLVG is included in Appendix A - Pilot survey.

4.4 Validating the questionnaire: factor analysis

The questionnaire will be validated utilizing factor analysis. Therefore, we conduct principal component analysis (PCA). By doing so, we aim to discover complex patterns to find potential factors (Yong, Pearce, et al., 2013). The factor analysis will prove whether all items for the constructs are statistically significant to classify the questionnaire as valid. The structure of the factor analysis is shown in fig. 4.3 and illustrates how each of the elements consists of multiple items. Each item stands for a perception statement within the questionnaire. Altogether, they are assumed to represent a theoretical construct. The factor analysis will demonstrate whether the statements indeed can be clustered within each of the presumed constructs.

Factor analysis aims to reduce measure variables into latent (unobservable) variables, sharing a common variance on a factor (Bartholomew, Knott, & Moustaki, 2011). The combination of items is evaluated through factor analysis to determine whether they are sufficiently coherent to combine them into an element. Hence, the factor declares the correlations by presenting the communality of items. Communality equals the common variance shared among a set of items, i.e. the higher they are correlated, the more variance they share (UCLA, 2021). On the contrary, unique variance represents the portion that is allocated to a specific item. Unique variance can be divided into two categories: particular variance of a specific item and error variance, which is anything unexplained (UCLA, 2021). An example of the accessibility factor is illustrated in fig. 4.3.

Figure 4.3: Itemization of theoretical constructs within questionnaire for factor analysis



A conventional factor analysis entails a mathematical model in which p indicates the number of variables $(X_1, X_2, \ldots, X_p)$, m indicates the number of underlying factors $(F_1, \quad F_2, \ldots, F_m)$ and X_j is proxy for latent factors (Yong et al., 2013). This results in the following equation:

$$X_j = a_{j1}F_1 + a_{j2}F_2 + \ldots\ldots a_{jm}F_m + e_j \tag{4.1}$$

$a_{j1}, a_{j2}, ..., a_{jm}$ are the factor loadings of a variable on the first factor. This basically indicates the amount a variable contributes to the factor (Harman, 1976). The factor analysis will be conducted in five main steps as explained hereafter:

1. **Step 1 - Coherence of items**
   The first step entails the verification of coherence between items. We do this by checking the communality of items. Items generally must exceed a communality of 0.25; a lower amount can be a reason to remove the item from the analysis. However, theoretically, well-founded items can be kept for research purposes even though communality of some items are lower.

After this step, it is no longer required to check the communality, as the factor loadings are more important further on

2. **Step 2 - Defining the factors**
   Factors with an initial eigenvalue of slightly above one are not considered solid factors. Each factor needs to contain at least two items, with a preferable number of three items. When there is merely one item with a high load on a factor, it takes no meaningful part in validating the factor analysis. In such cases, all items on a particular factor can be omitted, resulting in one factor less within the research. This exclusion has no consequences for the descriptive analyses in section 6.1.

3. **Step 3 - Simple structure**
   A simple structure requires each item to load high on one factor (i.e. higher than 0.50) and to load low on all other factors (i.e. lower than 0.30). This step can be conducted iteratively, in which variables can be excluded singly per step. There are some rules to approach the so-called simple structure based on the pattern matrix. Firstly, it is paramount to check how items load on each of the factors. The more they load on various factors, the more problematic it will be. Secondly, when multiple items have double loadings, the items with an approximate equal load can be excluded first. Third, items with a high factor loading on one factor and low on all others should retained. Lastly, variables with low factor loadings on all factors can be excluded. After variables are excluded based on factor loadings, it is no longer needed to check the communalities, as factor loadings are leading from here on.

4. **Step 4 - Orthogonality**
   The factor analysis starts with a crooked solution which is usually hard to interpret. As orthogonal solutions are easier to interpret, an orthogonal rotation will indicate how well it fits a solution. Hence, orthogonal solutions are preferred over crooked solutions. The decision for an orthogonal solution has two requisites. First, if the factor correlation matrix for the crooked solution displays high correlations, factors are usually not orthogonal. There are no complex rules in this, except for the boundaries mentioned earlier in the previous step. Second, we can compare the factor matrices of the orthogonal and crooked solutions. If the simple structure does not improve in the crooked solution, we may prefer an orthogonal solution.

5. **Step 5 - Interpreting factors**
   An interpretable solution requires items to load high on a factor to interpret it. If the items deviate too much, they still can be removed from a factor in this step. This decision requires to back to step 2. Hereafter, the factors can be interpreted with suitable labels. To conclude, the last step enables to validation of the relevant items on each of the factors. Moreover, it may confirm the validity of the survey.

4.5 Descriptive analysis

In section 4.2.2 we explained the reason the data could solely be used for descriptive analyses, given the limited sample size. However, descriptive statistics contain sufficient features to extract engaging lessons and compare the empirical and ethical domains. We serve two main lines of purposes with the descriptive analysis: conventional measurements provide insights on generic characteristics of the data, and relations between constructs will prove the discrepancy between the normative and descriptive.

4.5.1 Interpretation of results

The descriptive analysis consists of distinctive components to deduce patterns. As we show in fig. 4.4, we synthesise the literature, pilot survey and data from the survey to interpret the results. We only use the triangulation method in chapter 6, as it enables to evaluate all sub-parts of this study entirely.

4.5.2 Measurements

Descriptive statistics entails the collection, description and interpretation of data to formulate conclusions about a sample (Pérez-Vicente & Ruiz, 2009). Qualitative (e.g. organisation, job period, job type, etc.) and quantitative variables (Likert scale of statements) are included. Both

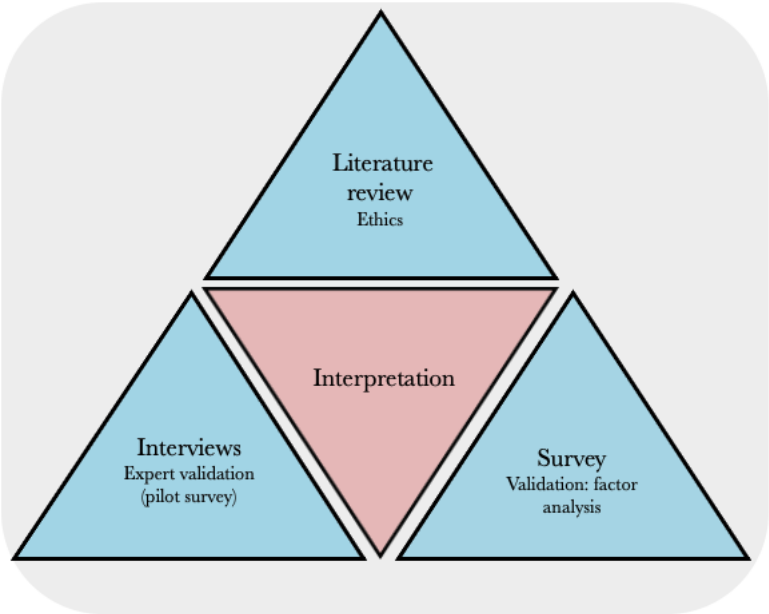


Figure 4.4: Triangulation method

qualitative and quantitative measurements require statistical methods. We elaborate on both types hereafter.

Qualitative variables can be described numerically by showing the absolute frequencies, relative frequencies, percentages and rates. The methods support the understanding of the personal characteristics of the sample and provide insights on the relative proportions between respondents from various organisations and job types. Moreover, it illuminates the implications of the conclusions if there are any imbalances between certain categorical groups.

Descriptive statistics with quantitative variables can be split up into three main categories: dispersion, shape and position measurements (Pérez-Vicente & Ruiz, 2009). Position measurements are used to compute the mean and median of certain variables (e.g. mean of one statement or entire theoretical construct). The mean will be calculated with the following basic formula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.2}$$

Dispersion measurements indicate the variance of the data set, which will be calculated accordingly:

$$s_{n-1}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{4.3}$$

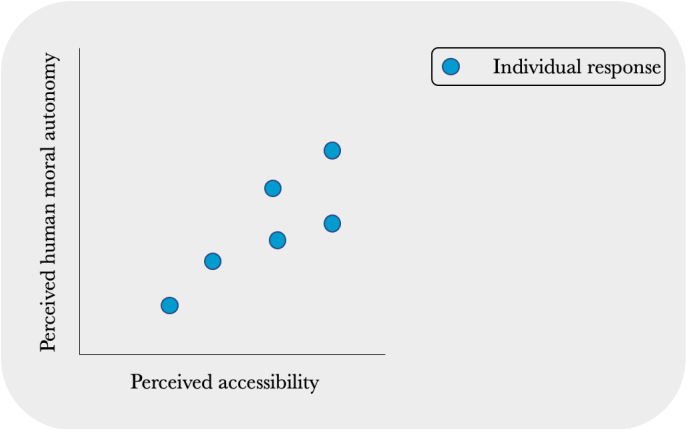
The standard deviation can be calculated by taking the square root of the variance. The standard deviation is used to express the dispersion of the mean value of a variable.

The shape measurement is used to measure the symmetry of the variable to show the distribution of the mean (Pérez-Vicente & Ruiz, 2009). The skewness of a variable can result in three types of values: zero, negative or positive. Zero suggests the sample values are distributed. Equally, negative skewness indicates concentration on the right of the mean (left-skewed) and vice versa for positive values (right-skewed). Lastly, data points with an unusual distance from the overall sample may occur, which we define as outliers. Outliers are defined whenever they take place out of the range, computed as follows: $[\mathrm{P}_{25} - 1.5\mathrm{QR}, \mathrm{P}_{75} + 1.5\mathrm{QR}]$.

4.5.3 Relationships and comparisons

The actual synthesis of the normative and descriptive will take place in this part of the study. We do this by plotting the theoretical constructs as a function and putting respondents' perceptions along the axes. An example is given in fig. 4.5, in which the red line indicates the perception of Van den Hoven (1998) on the relationship between accessibility and HMA. The blue dots are the perceived theoretical constructs as answered by the respondents. We can visually express the comparison between how philosophers ideally define their vision and how respondents perceive BAIT along with the defined theoretical constructs.

Figure 4.5: Example of descriptive analysis



Summary box chapter 4

The goal of chapter 4 is to explain the methodology this research entails. The following points summarise this chapter:

- We use a mixed-method approach to construct the HMA survey. This approach consists of a qualitative, quantitative and results part.
- We perform the interviews to pre-test the survey. This is a first validation step for the survey to preserve its validity.
- The survey is designed to conduct factor analysis. This is the second validation step. We included four perception statements to measure each of the hypothesised constructs. Only one importance statement per construct is included
- The survey results enable to perform a descriptive analysis to find exciting relationships between constructs.
- We make use of the triangulation method to analyse the descriptive results. This method entails a literature review to best practices, the pilot survey and results from the survey. Altogether, the results of the sub-parts are interpreted.

Chapter 5

Results: The HMA Survey

> "A free man is one who lives under the guidance of reason, who is not led by [emotion] ... but who directly desires that which is good"
>
> Spinoza, Ethica, p.232

This chapter aims to (1) present a measurement instrument to measure perceptions of HMA in BAIT and (2) validate the survey. In section 5.1 the context regarding the respondents and the organisations are clarified. Thereafter, in section 5.2 we describe the main results of the pilot survey. This forms the pre-validation of the survey. In section 5.3 the survey structure is explained. In section 5.4 the survey itself is presented and explicated. Thereafter, in section 5.5 we briefly explain the validation. The latter also addresses the implications of the survey. To conclude, this chapter aims to answer the following sub-question:

Research question 4

How can we measure the degree to which DSSs like BAIT respect human moral autonomy using a questionnaire?

5.1 Respondents description

This section describes the organisations, the respondents and provides additional context. The organisations participated with the problem owner, Councyl, to optimise a specific decision-making within their field of expertise.

5.1.1 Deloitte experts

The first organisation that worked with BAIT through the pilot of Councyl is Deloitte. Deloitte has 5500 employees spread out over 14 offices throughout the Netherlands, providing professional services in the domains of accountancy, tax, consultancy, risk and financial services (*Deloitte*, 2021). Deloitte fulfils its prominent role in the public and private sector and possesses expertise in almost all disciplines. Hence, many decisions are made each day to provide their clients with customised advice for improved strategy. The cooperation with Councyl was set up to improve their decision-making regarding public tenders. The scope of Deloitte entails the qualification of public tenders. This includes the design and presentation of the choice experiment. Accordingly, the model will be predicted containing weights for criteria to represent the possible trade-off of users. After that, an introspective report is written and presented to reveal the decision criteria.

5.1.2 OLVG experts

The second organization that aimed to improve its decision-making is OLVG. OLVG is a hospital attached to the medical centre of AUMC, which is associated with the University of Amsterdam. OLVG has a capacity of approximately 555 beds, it dealt with over 48,000 emergency admissions

and is specialized in cardiology, HIV treatments, toxicology, orthopaedics, first aid, and intensive care (IC) (*OLVG*, 2021). The IC department is concerned with seriously ill patients, including COVID-19 patients. Decision-making within this department is of great interest since the COVID-19 pandemic. Especially since the capacity has been challenged throughout the period. On that account, OLVG was interested in their decision-making regarding the admission of patients to the IC. Their objective is two-fold: firstly, they want to obtain insights on the trade-offs and secondly to support efficient and qualitative decisions. For OLVG, the knowledge of various intensivists is modelled to predict decisions regarding the admission of COVID-19 patients to the IC. The primary purpose of this cooperation is to test the method. Moreover, it aims to retrospectively and prospectively validate the model.

5.1.3 UMCG experts

The last included organisation is the UMCG. The total capacity of this medical centre is approximately 1330 beds and has over 12,000 employees (*UMCG*, 2021). Councyl co-operated with the Department of Surgery, Division of Pediatric Surgery to design a DSS for one specific decision-making. The case of UMCG is scientifically reported by ten Broeke et al. (2021). It illustrates the operationalisation of BAIT to decide on operating patients with a premature Neonate with pre–necrotising enterocolitis (NEC). UMCG operates within the same sector as OLVG and therefore faces similar medical ethical issues. The project of UMCG includes the qualification of criteria to serve patients on a premature neonate with NEC. For this case, too, the objective was to illustrate the functionalities and applications of BAIT. Hence, the goal was not to present the surgeons with new insights but rather to design and validate the model within their decision-making.

5.2 Pilot survey

The pilot survey is conducted with a group of decision-makers from two out of three participating organizations. The semi-structured interviews are performed to pretest the survey on intelligibility and identify differences in perceptions. The interviews lasted for 30-45 minutes, in which the experts are invited to explain their feelings regarding the technology. Moreover, they are asked to describe their perception of some of the theoretical constructs. section 8.3 contains the complete interviews. This section entails a brief analysis of both groups.

Deloitte set up a decision model with Councyl to optimize their decision-making on public tenders. They appreciated the added value BAIT might provide for their decision-making. More importantly, they value the learning process as they were asked to quantify essential criteria. Some criteria they did not consciously think of before, resulting in further understanding in the development phase of the model. However, the experts on this particular decision-making do seem to be worrying about using such DSS regularly. They would primarily be worried when used for automation purposes instead of a supportive tool. To develop a so-called 'industry sensitivity, they argue, every expert should make the decisions without such models. They did, nevertheless, get some additional insights with regards to human expertise. Based on the first results of the model, they argued there is the potential ability to identify opportunism in the decision-making among their experts. By doing so, they could track the extent to which the decision-makers make their decisions 'rationally' (i.e. based on the determined criteria instead of gut feelings). To conclude, the interviewees from Deloitte thoroughly thought about how BAIT could potentially be used within their organization and what consequences it may occur.

What we observe in the answers given by the interviewees of Deloitte is that they accurately did understand the theoretical constructs of HMA in the usage of BAIT. It was even more intuitive to the interviewees whenever they were asked how it potentially could be used and what consequences it may cause. Overall, they did seem to be familiar with the terminology. Hence, the potential respondents of Deloitte are likely able to provide exciting answers to the survey to study their perceptions on HMA in the usage of BAIT.

OLVG set up a decision model with Councyl to optimise their decision-making on the admission of patients to the IC of the hospital. The experts were impressed with the predictive values, mainly because they did not define the criteria for this decision as before. The exciting point within this interview is that the intensivist did not seem to feel any 'threat' regarding his moral autonomy when such models are used on the work floor. He emphasised multiple times the importance

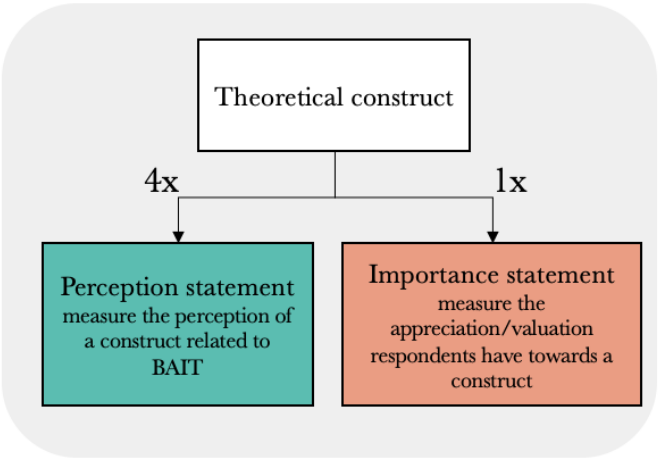
of using it solely as a supportive tool to gain insights. The experts in his department always bear moral responsibility for their decisions concerning their patients. This notion is present and seems to play an essential role in making a decision. Moreover, the interviewee appeared to be highly conscious of specific disclaimers of the model when integrated into the decision-making. BAIT captures like/dislike of colleagues (KPIs are not defined in performance terms). It contains randomness (noise). It is always essential to take external factors into account before making a decision. The model has not been integrated within the organisation of OLVG (yet). However, this interview gave the impression that the concept of BAIT is sufficiently understood to participate in this survey.

The interviewee of OLVG is not only conscious of how BAIT could be integrated within the organisation or this particular decision-making. Moreover, he understands how it may affect his moral autonomy. He related technical features of the technology with decision-making and was able to sum potential consequences. Hence, he (and his colleagues) seemed to be highly proficient at providing answers on the effects on HMA of potential BAIT users.

5.3 Questionnaire structure

The theoretical constructs are each subdivided into two separable statements: importance and perception. The importance statements enable us to review the valuation of respondents towards the construct. The perception statements measure the experience respondents have within the use of BAIT. For each of the theoretical constructs, one importance statement is included and four perception statements (see fig. 7.2). Both can be used for descriptive analysis, but the latter contains more statements to perform factor analysis. By doing so, we measure two distinctive types: a respondent may rate a construct highly through the interest statement (e.g. that accessibility of the system is indeed paramount as the underlying theory argues) but could simultaneously disagree that BAIT meets the conditions that support the construct. This inquiry could lead to exciting conclusions. Respondents can rate a construct high on the importance statement but low on perception and vice versa.

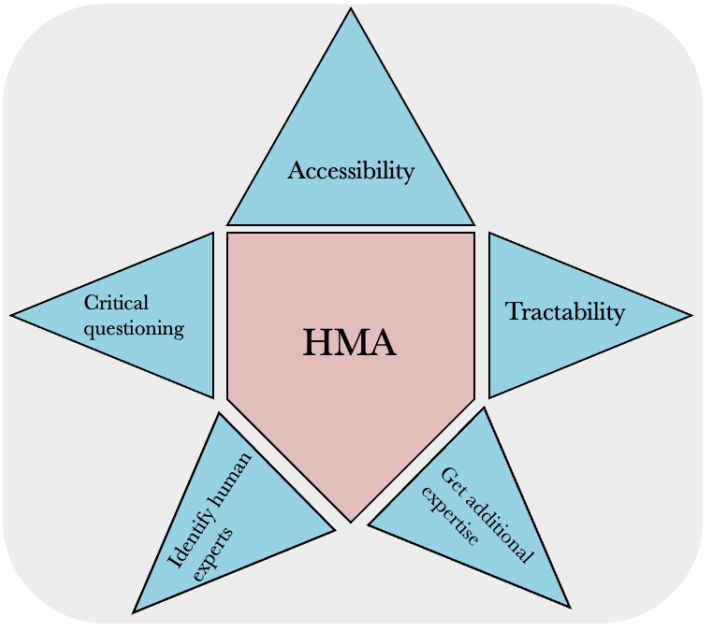
Figure 5.1: Types of statements



Following the line of van der Waa et al. (2021), we used implicit measurements to assess the perceptions of respondents on each of the constructs. This methodology aims to reveal the discrepancies between ethical acceptability (i.e. the view of philosophers) and social acceptance (i.e. the respondents' opinion). Therefore, appropriate, unambiguous constructs are compared across the two ethical arenas. We do this by using importance statements, which reveal the importance respondents ascribe to each moral construct. This directly clarifies the relation between social acceptance and ethical acceptability. In contrast, the perception statements reveal the experience of respondents within a particular technology. Those statements measure the social acceptance of a specific technology. Hence, the perception statements reveal the relation between social acceptance and ethical acceptability indirectly.

For practical reasons, we retrieve only a limited number of constructs from the theoretical framework to the empirical part of this study. We include the following constructs: (1) HMA, (2) accessibility, (3) tractability, (4) get additional human high-fidelity expertise, (5) identify & empower human experts and (6) critical questioning of decision-making by a human agent. The selection of constructs is first of all defined based on applicability. Some of the constructs are less applicable and measurable (e.g. fix bugs and programming errors) than others (e.g. perception on accessibility). Moreover, the selection of this particular combination addresses the potential beneficial features of BAIT that Councyl has claimed. BAIT is often framed as an interpretable, transparent model (covered by scrutinizability condition). Moreover, BAIT aims to solely support experts' decisions (covered by the "getting human expertise" construct). It focuses on providing insights into decisions to decision-makers (critical questioning and evaluation condition). Lastly, measuring this system's perceived HMA of respondents is crucial, which is captured in the overarching HMA construct. To conclude, the empirical part will only include those constructs.

Figure 5.2: Selection theoretical constructs



5.4 Measurement instrument: HMA survey

We formulated and clustered the statements based on the theoretical constructs. As explained before, the statements are of two types: importance and perception. For both types, the aim is to measure the construct implicitly. The statements within the survey are listed in table 5.2.

The survey is constructed to measure the concerned constructs as those are deemed necessary in the context of HMA. The first block in the actual survey (see appendix section 8.3) contains all importance statements, which are presented to the respondents to understand the valuation they assign to each of the constructs. This distinction enables researchers to compare what respondents value and how respondents perceive the constructs within a technology context. The importance statements all end with the label "_int" (see table 5.2). Hence, the importance statements can easily be distinguished in further analysis from the perception statements. Furthermore, each construct contains four perception statements to measure the experience respondents have regarding BAIT as a technology. As explained before, for each construct, four statements are formulated to perform a PCA. The perception statements are logically coded with the abbreviation of the concerned construct, from one to four, respectively. The underlying assumption is that the statements coherently define each of the constructs. The validation in the following subsection will prove whether this is true.

Table 5.1: Items in HMA survey

cyan **Construct**	**Statement type**	**Statement**	**Code**
HMA	Importance	I think it is important not only to depend on the Councyl model when making choices	MA_int
Accessibility	Importance	I think it is important to have access to all forms of information that the Councyl model offers	ACC_int
Tractability	Importance	I think it is important that I have insight into the weight of each factor that is included in the Councyl model	TRAC_int
Getting human high-fidelity expertise	Importance	I think it is important to always have the option to consult colleagues in my department when making choices	GET_int
Identify & empower human expertise	Importance	I think it is important that the Councyl model stimulates the development of expertise in me and my colleagues	EXP_int
Critical questioning and evaluation	Importance	I think it is important that I can critically and independently assess the development of the advice of the Councyl model	CRIT_int
HMA	Perception	1 - If I use the Councyl model, I can still take responsibility for my choices	MA1
HMA	Perception	2 - When I use the Councyl model, I feel limited in making my choice	MA2
HMA	Perception	3 - I consider the advice of the Councyl model to be non-binding	MA3
HMA	Perception	4 - Using the Councyl model does not diminish the moral responsibility I bear for my choices	MA4
Accessibility	Perception	5 - The Councyl model offers a sufficient degree of accessibility to relevant information	ACC1
Accessibility	Perception	6 - The Councyl model allows me to understand how the system works	ACC2
Accessibility	Perception	7 - The Councyl model offers the correct and relevant information for making choices	ACC3
Accessibility	Perception	8 - Information about the development of the advice is clearly presented	ACC4
Tractability	Perception	9 - The Councyl model enables me to study the development of the advice properly	TRAC1
Tractability	Perception	10 - The Councyl model enables me to sufficiently analyze the advice	TRAC2
Tractability	Perception	11 - The Councyl model offers me the opportunity to trace back the advice on the basis of criteria and weights	TRAC3
Tractability	Perception	12 - The Councyl model enables me to trace the formation of the advice	TRAC4
Getting human high-fidelity expertise	Perception	13 - The Councyl model offers me the opportunity to consult colleagues when making choices	GET1
Getting human high-fidelity expertise	Perception	14 - Using the Councyl model encourages me less to approach my colleagues for additional advice	GET2
Getting human high-fidelity expertise	Perception	15 - The Councyl model leads to an improvement of the dialogue between me and my colleagues	GET3
Getting human high-fidelity expertise	Perception	16 - The Councyl model stimulates the exchange of ideas within my team for making choices	GET4
Identify & empower human expertise	Perception	17 - The Councyl model leads to an improvement of competences within my team	EXP1
Identify & empower human expertise	Perception	18 - Using the Councyl model makes it more difficult to recognize (new) experts in my department	EXP2
Identify & empower human expertise	Perception	19 - The use of the Councyl model leads to an improvement in my professional knowledge	EXP3
Identify & empower human expertise	Perception	20 - The Councyl model does not stand in the way of the development of competences of colleagues of mine and my colleagues	EXP4
Critical questioning and evaluation	Perception	21 - The Councyl model enables me to critically reflect on my choices	CRIT1
Critical questioning and evaluation	Perception	22 - When I use the Councyl model, I can still critically consider my choices	CRIT2
Critical questioning and evaluation	Perception	23 - The Councyl model provides the appropriate information to be able to critically execute decision-making	CRIT3
Critical questioning and evaluation	Perception	24 - The Councyl model provides sufficient information to be able to critically assess the advice	CRIT4

5.5 Validation of questionnaire

We designed the survey to measure all of the concerned theoretical constructs. However, a factor analysis demonstrates the validity of the assumed underlying constructs within the survey. This section elucidates the principal component analysis (PCA) of the survey results as shown in fig. 5.3.

We explored all 24 items of the measurement instrument in the PCA to examine the factorial structure of the HMA survey. Hence, we included only the perception statements. The Kaiser-Meyer-Olkin measure did not verify the sampling adequacy for the analysis, as well as the Bartlett's test of Sphericity. Both measurements indicate the correlation structure is not adequate for factor analysis. This inadequacy was already assumed beforehand, as the sample size remained low (N = 11). However, given the current input by the limited number of respondents, we can carry on the PCA. Yet, this has implications for the results of this study, as no conclusive evidence can be declared through this validation.

The PCA is conducted with a six-factor solution, as six hypothesized constructs are included initially. This solution resulted in an explained variance of 86.04%. Unsurprisingly, none of the six factors obtained had the same structure as the underlying constructs imply. Hence, we have to rephrase the components based on the output of the rotated component matrix. Only a limited number of items are selected for each component to form the basis of the regarding factor. All included items possess a high factor loading within the cluster. Therefore, we excluded all other items. Hence, another PCA is conducted with only the relevant items. The final PCA resulted in an explained variance of 91.56%. To conclude, we validated the survey by adjusting the components and the underlying meanings.

Within the first component, statements 21, 8 and 22 are selected. The two statements regarding critical questioning and evaluation show coherence with the statement on a clear presentation given by the model. The name of this component is 'Choice reflection'. Even though distinct constructs are combined here, they do measure a vital aspect of HMA. Given the formulation of the statements and the initial constructs they represent, we decided to rephrase this construct as: "Critically reflect the presented information". Hence, this component aims to measure the extent to which respondents believe they can critically assess the decision-making.

The second component includes statements 20, 2 and 5. Those components originate from three different constructs. Statement no. 20 mentions the restriction of human expertise, statement no. 2 mentions the limitation in making choices, and statement no. 5 mentions the accessibility to relevant information. The name of this component is 'Autonomy'. The shared denominator of the statements is the general notion of restriction. Therefore, we rephrase this component as "Restriction by the technology". This component aims to measure the extent to which respondents feel restricted in decision-making using the technology.

The third statement includes statements 24, 6 and 23, derived from two distinctive constructs. Similarly to the first component, this includes the critical questioning and accessibility constructs. The name of this component is 'Boundless decision-making'. The accessibility statement in the first component appoints the presentation of information. On the contrary, the accessibility statement in this component mentions how the respondent understands the system. Hence, this component is now defined as "Critical understanding the technology". This component measures the extent to which the respondents perceive their critical understanding of the concerned technology.

The fourth component includes statements 1 and 4. Both originate from the HMA construct. Hence, we decide to maintain this construct through those statements. The name of this component is 'Intelligibility'. The fourth component, therefore, is defined as "Human moral autonomy". This component measures the extent to which respondents feel autonomous in making their decision in a technology context.

The fifth component includes statements 9 and 3. The statements represent the tractability and HMA constructs, respectively. The name of this component is 'Explainability'. Whereas the former statement mentions the extent to which the user can study the advice of the technology, the latter appoints the user's autonomy. Collectively, this component is rephrased as "Understand the

Table 5.2: Final constructs HMA survey after validation

Component name	Perception statements	Definition	Initial constructs	Cronbach's Alpha
Choice reflection	8, 21, 22	Critically reflect the presented information	Critical questioning, accessibility	0.874
Autonomy	1, 4	Human moral autonomy	HMA	0.587
Boundless decision-making	2, 5, 20	Restriction by the technology	Identify expertise, HMA, accessibility	0.775
Intelligibility	3, 9	Understand the technology and my role	Tractability, HMA	0.904
Explainability	6, 23 ,24	Critical understanding the technology	Accessibility, critical questioning	0.789
Knowledge interchange	7, 15	Improved dialogue on decision-making	Accessibility, get expertise	0.583

technology and my role". The component aims to measure how the respondent can study advice and not feel obliged to follow up on the advice given by technology.

The last component includes statements 15 and 7. The statements represent the accessibility and get expertise constructs, respectively. The name of this component is 'Knowledge interchange'. The former statement mentions the interaction between human experts, whereas the latter appoints the correctness of information provided by the technology. This component is rephrased as "Improved dialogue on decision-making". The component aims to measure how respondents perceive improvement of the dialogue between peers through technology.

We could not maintain the initial structure of the survey with the validation. However, we were able to produce a survey that is practical and accurate. Each of the components still reveals essential aspects of HMA in the context of technology. Hence, based on this first validation, researchers can use the validated version of the survey to study some of the critical aspects of HMA. By doing so, perceptions of HMA within the context of specific technologies can be measured and evaluated.

Figure 5.3: Final PCA

Rotated Component Matrix[a]

	Component							
	1	2	3	4	5	6	Construct	Original construct
8 - Information about the development of the advice is clearly presented	0,925		-0,274			-0,160	Construct A	Accessibility
22 - When I use the Councyl model, I can still critically consider my choices	0,844	0,210		0,249				CRIT
21 - The Councyl model enables me to critically reflect on my choices	0,836			0,311	0,133			CRIT
4 - Using the Councyl model does not diminish the moral responsibility I bear for my choices	0,304		0,179	0,917			Construct B	HMA
1 - If I use the Councyl model, I can still take responsibility for my choices	0,286	0,220		0,859	0,222			HMA
20 - The Councyl model does not stand in the way of the development of competences of colleagues of mine and my colleagues	0,321	0,904			0,180		Construct C	Identify expertise
2 - When I use the Councyl model, I feel limited in making my choice	0,236	0,892	0,116	0,270	0,180			HMA
5 - The Councyl model offers a sufficient degree of accessibility to relevant information	-0,457	0,794	-0,143	-0,101		0,339		Accessibility
9 - The Councyl model enables me to study the development of the advice properly	-0,130	0,134	0,193	-0,108	0,929		Construct D	Tractability
3 - I consider the advice of the Councyl model to be non-binding	0,252	0,295	0,134	0,367	0,812			HMA
6 - The Councyl model allows me to understand how the system works	-0,205	-0,136	0,889	0,139	0,101	-0,205	Construct E	Accessibility
24 - The Councyl model provides sufficient information to be able to critically assess the advice			0,821	0,122		-0,146		CRIT
23 - The Councyl model provides the appropriate information to be able to critically execute decision-making		0,239	0,793	-0,226	0,159	0,284		CRIT
15 - The Councyl model leads to an improvement of the dialogue betweenme and my colleagues		0,171	-0,148	0,111	0,314	0,877	Construct F	Get expertise
7 - The Councyl model offers the correct and relevant information for making choices	-0,263			-0,250	-0,473	0,774		Accessibility

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Summary box chapter 6

The goal of chapter 6 is to present the results of this study: the HMA survey and the validation utilizing factor analysis. We did this by applying the methodology as explained in chapter 4. The survey contains six hypothetical constructs; each construct contains four perception statements and one importance statement. Through a PCA, we were able to adjust the survey and validate it partially. The following points summarise this chapter:

- The interviews with some respondents indicated they were overall proficient at filling in the survey. Moreover, they provided valuable information on their perceived moral autonomy. Lastly, they were fully aware of how the DSS could be integrated into the decision-making.
- Three organisations participated in our study: Deloitte, OLVG and UMCG.
- The HMA survey is directly built on the HMA framework. For feasibility reasons, we selected a limited number of hypothetical constructs in the methodology.
- The HMA survey consists of six hypothesised constructs from the HMA framework: (1) HMA, (2) accessibility, (3) tractability, (4) get high-fidelity human expertise, (5) identify & empower human expertise and (6) critical questioning and evaluation.
- The survey contains two types of statements: perception statements and importance statements. The former measures the perception of a construct related to BAIT, whereas the latter measures the valuation of a construct.
- The number of respondents appeared too low to validate the correlation structure.
- As we continued the validation, we set up new components, consisting of statements from various constructs.
- The new components are: (1) Choice reflection, (2) Autonomy, (3) Boundless decision-making, (4) Intelligibility, (5) Explainability (6) Knowledge interchange.

Chapter 6

Improving HMA of BAIT: analysis and advice

> "The human being's autonomous will causes its totally free action (which can be moral and, therefore, can be punished if it fails)."
>
> Critique of practical reason, Kant 1788: 5, 100

This chapter entails the descriptive analysis and an advisory guideline to improve BAIT in terms of HMA. Hence, it is an application and evaluation of the measurement instrument. The descriptive analysis of the data demonstrates how insights can be obtained from the HMA survey. For this demonstration, similar data as in the survey validation is used (see section 5.5). By doing so, we aim to provide relevant insights for the problem owner. Additionally, for each theoretical construct, the information from the literature is synthesized with the perception of experts. The synthesis is a tangible way to bridge the gap between social acceptance and ethical acceptability. Effectively, we aim to formulate an answer to the following sub-questions:

Research question 4 & 5

4. To what extent does BAIT respect the conditions of HMA according to end users?
   (a) What importance do end users ascribe to the philosophical conditions of HMA?
   (b) Does BAIT respect the philosophical conditions of HMA in the perception of end users?
5. How can DSSs like BAIT be designed so that they better respect human moral autonomy?

In section 5.2 the results of the pilot survey are explained. This section clarifies the interviews to pre-test the survey. It contains interesting insights into the opinion of decision-makers regarding BAIT. After that, in section 6.1 the entire descriptive analysis is outlined, showing insightful visualisations. The guidelines, which entails valuable findings from the literature, is written in section 6.2.

6.1 Descriptive analysis

This section presents the descriptive results of the survey. It shows insightful visualisations on sample characteristics, a high-level analysis, distribution of scores and the relationships between various hypothesised constructs.

The descriptive analysis demonstrates the potential ability the HMA survey possesses to measure perceptions. Hence, it forms the visual presentation of the input and is used for learning

and development purposes. By doing so, we aim to obtain information on the perceptions of respondents on the hypothesised constructs. To conclude, the HMA survey can be used for a more thorough analysis to understand the empirical results concerning normative concepts.

6.1.1 Sample information

This subsection provides information on some of the sample characteristics. The distribution of the survey involves three organizations: Deloitte, OLVG and UMCG. The number of responses for each organization is presented in fig. 6.1. To reach all potential respondents, the experts who participated in the former choice experiments to construct the DSS in collaboration with Councyl were invited to fill in the survey. Approximately 25% of the invited experts have filled in the survey. Moreover, the analysis makes a distinction between job types among the respondents; fig. 6.2 presents this. The two graphs are merely used to provide insights on the personal characteristics of the respondents and are not used for further analysis. However, they do show a notable discrepancy in the number of respondents between UMCG and the others. On that account, we also observe a balanced distribution of job types from the respondents of UMCG.

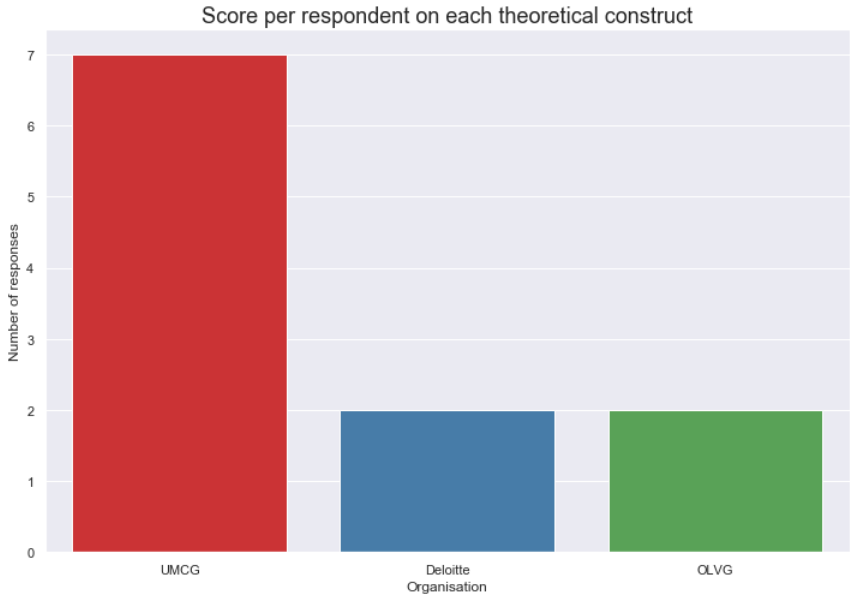


Figure 6.1: Number of respondents per organisation

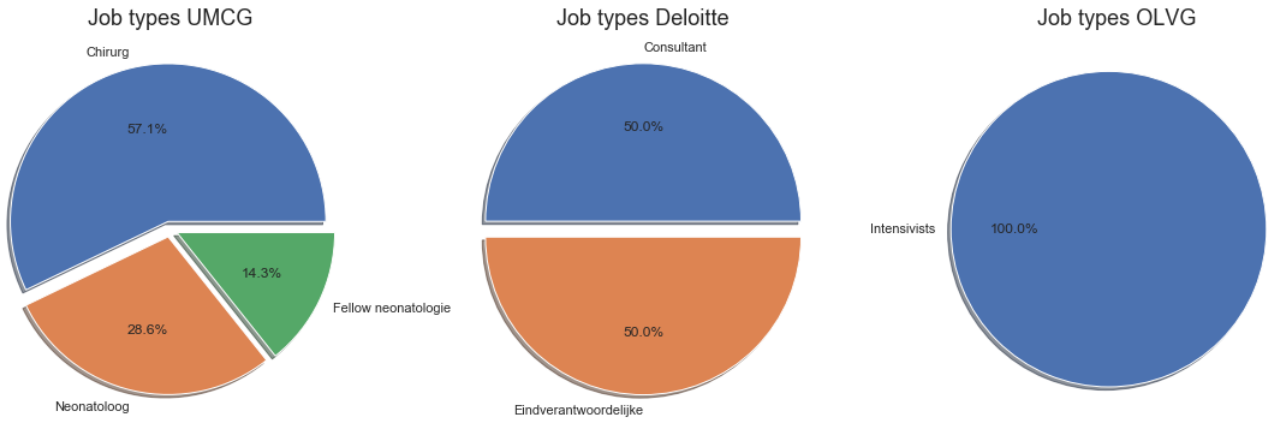


Figure 6.2: Share job type per organisation

6.1.2 High-level analysis

This subsection presents the respondents' answers to the statements formulated to measure the constructs of HMA. The proposed statements are rated on a Likert scale. The scale varies in the following levels: 1=Strongly disagree, 2=Disagree 3= Neutral, 4=Agree, 5=Strongly agree. The average rating per statement is calculated by summing up the scores of all respondents and dividing the value by the sample size. fig. 6.3 and fig. 6.4 (perception and importance respectively) show the mean, standard deviation, min/max values and percentile indicators in the form of boxplots.

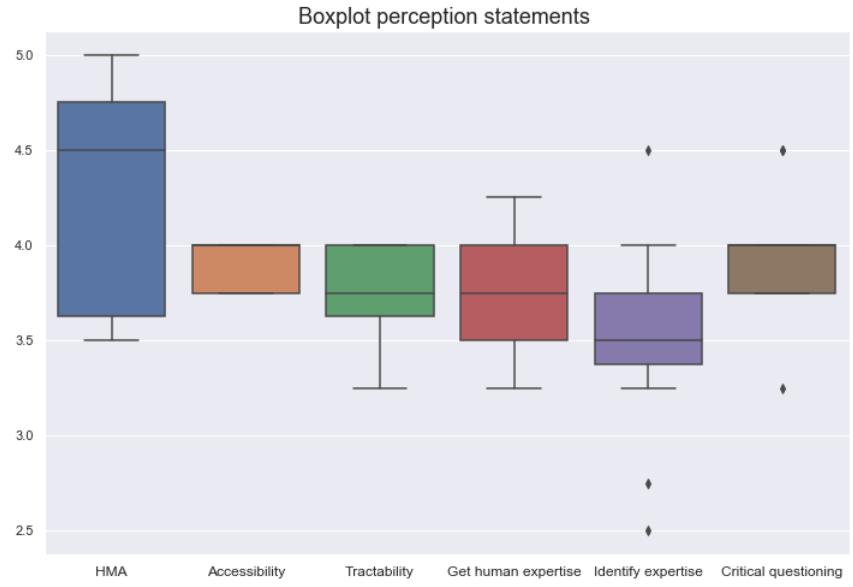


Figure 6.3: Boxplot of perception statements for all constructs

Figure 6.4: Boxplot of importance statements for all constructs

The results in both figures show that respondents predominantly agree with the proposed statements. This result indicates decision-makers perceive BAIT to meet the theoretical constructs

quite sufficiently. However, it also becomes clear that identifying and empowering expertise scores lowest on average. This notion was explicitly mentioned in the interviews for the pilot survey (see section 5.2. In fig. 6.5 we can observe that within that construct, the statement "Using the Councyl model makes it more difficult to recognise (new) experts in my department" scores lowest. The scores on this statement are inverted, as it is negatively formulated. Hence, most respondents do agree with the statement that it would become harder to identify new experts. This finding might be, what the experts at Deloitte referred to earlier in the interviews, the ability to develop an "industry sensitivity". With a mean value of 3.36, statement EXP1 also scored relatively low. This statement addresses the potential improvement of human competence by BAIT. The respondents are somewhat less convinced that BAIT supports human competency in comparison to the other statements. The same heatmap in fig. 6.5 is also clustered based on the dendrogram methodology, presented in section 8.3. With the euclidean metric, we found out that the respondents could be clustered based on the organisation.

The correlations between statements within constructs are displayed in fig. 6.6. Let us refer to this figure as intraclass correlations, in which the classes are the theoretical constructs. We can observe that most statements are either positively or negatively correlated, but none of them indicates a strong correlation (maximum value of 0.30). However, within the tractability and critical questioning and evaluation constructs, we see correlations close to zero. Strong correlations are required to validate the survey in the future. The correlation matrix between all of the statements and constructs are included in section 8.3.

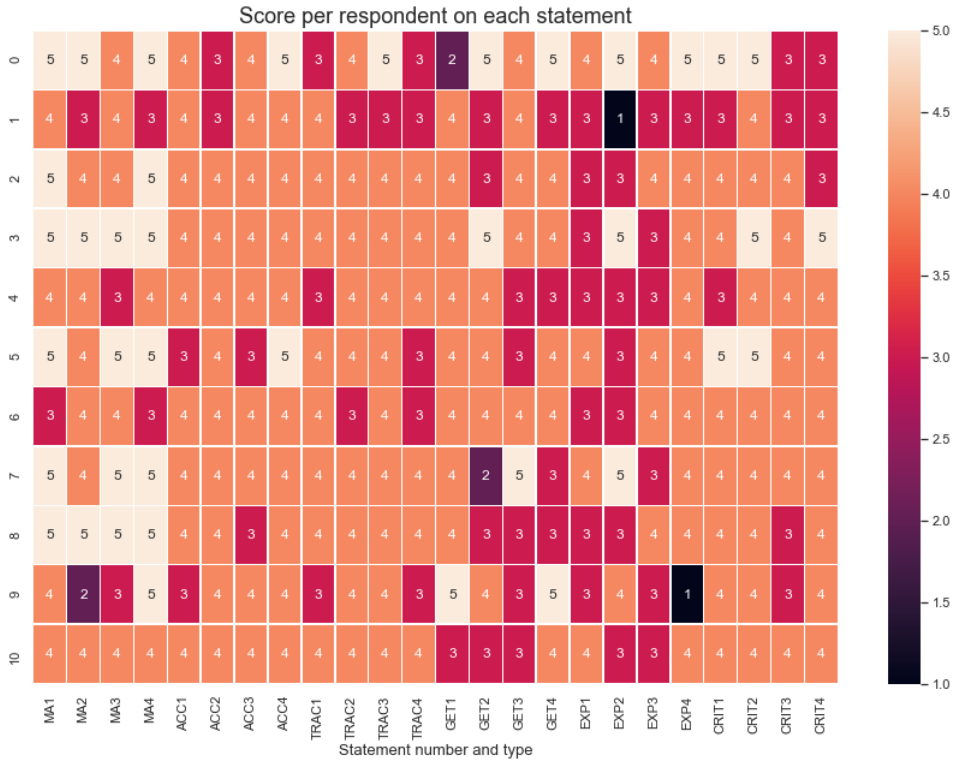


Figure 6.5: Scores by all respondents on each statement

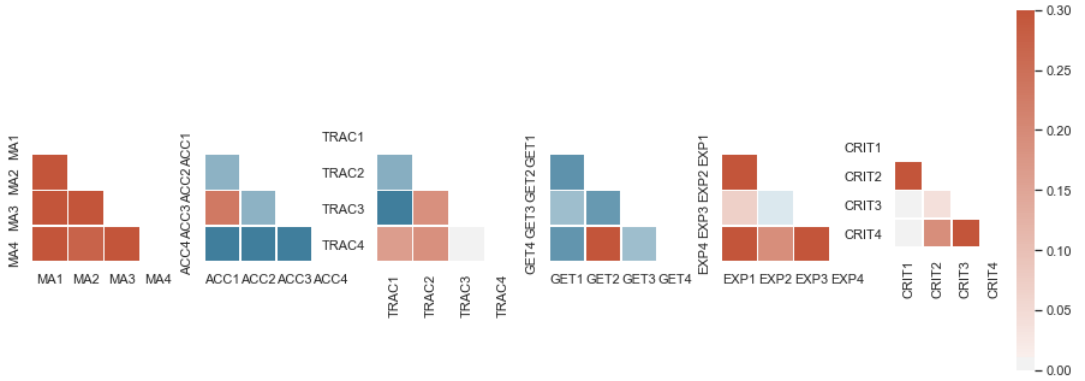


Figure 6.6: Correlations within constructs

6.1.3 Distributions of scores

This subsection presents the results on the distribution of scores given by respondents. The distribution scores enable us to study the difference between the perception of people of a particular technology and the importance they ascribe. A non-parametric test (t-test) is conducted for each comparison to determine the significance of the difference between both measures.

The distribution of scores gives a good indication of the dispersion of data, as it is demonstrated in fig. 6.7. The coloured lines in both graphs indicate the weighted average of the distribution. The black lines indicate a normalized distribution of the data. As argued in the previous subsection, respondents predominantly agree with the statements on HMA of BAIT. The left-hand of the figure implies that the mean value among all statements on perceptions is approximately 3.85. On the right-hand-side we observe the mean importance is around 4.4 for all constructs. This value implies the constructs are deemed imperative by the respondents. However, the disclaimer must be made that the constructs to measure the importance only contain one statement instead of the perception constructs in which four statements are included. The first results indicate the perception of these constructs of HMA is scored slightly lower than the actual importance respondents ascribe. Through a t-test, we proved that the difference between the mean of both measures is statistically significant, as shown in fig. 6.8. The distribution of scores for each construct is included in section 8.3 to provide a more profound understanding of the dispersion of scores. Moreover, a similar significance test is conducted for each of the comparisons. To conclude, a statistical comparison between both measures provides valuable information to find discrepancies. By doing so, developers can identify the most critical features of the technology, according to respondents.

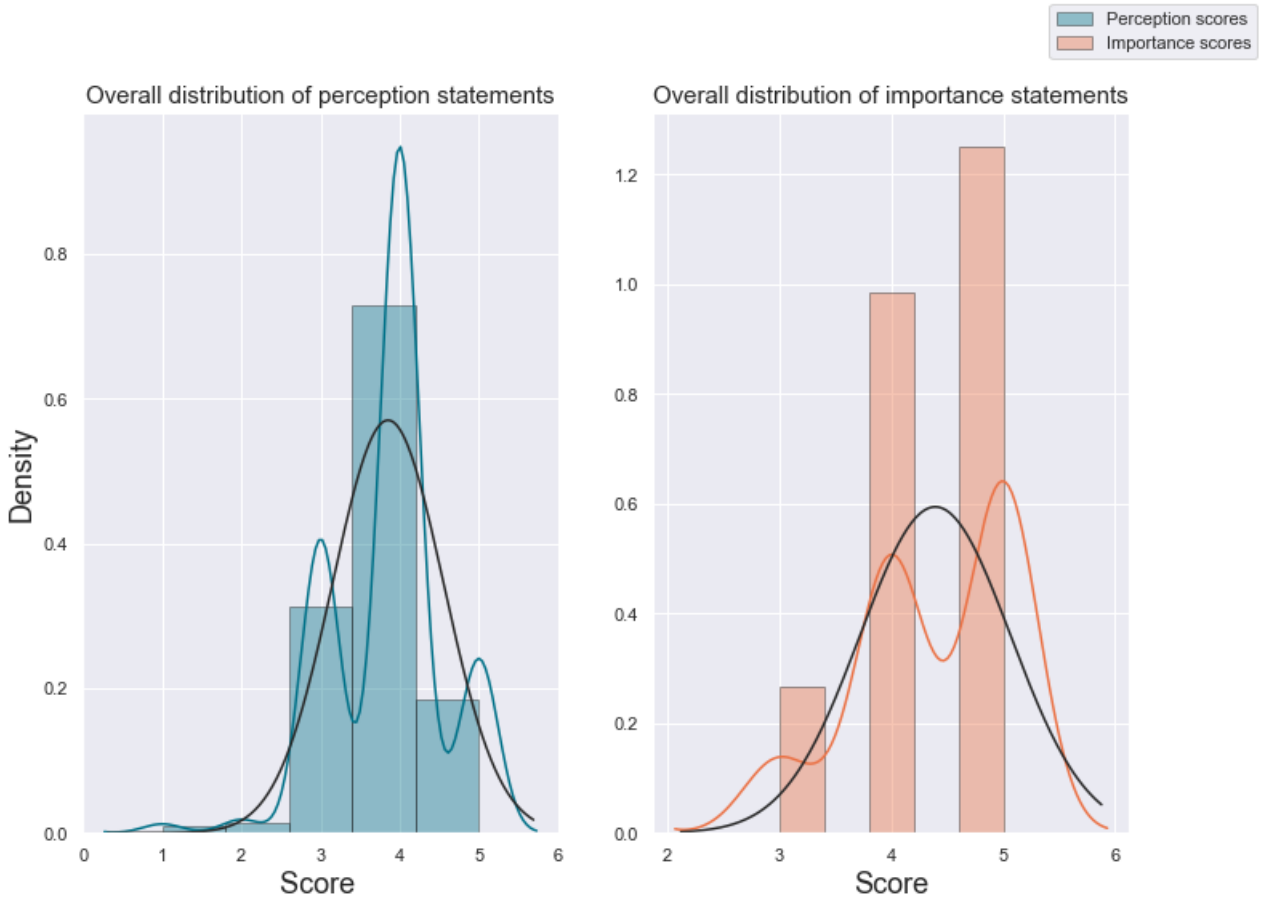


Figure 6.7: Distribution of all scores. Left: perception score distribution, Right: importance score distribution

	Test statistic	p-value
Sample data	-3.570317030082882	0.005093057103273934

Figure 6.8: t-test of difference between perception and importance statements

6.1.4 Relationships between constructs

The last subsection entails the visualisation of relationships between the HMA construct and all other constructs. This is done for both perception statements (blue scatterplot, see fig. 6.9) and importance statements (orange scatterplot, see fig. 6.10). By doing so, we enable ourselves to identify the supposed relationships as determined in chapter 3. The theoretical framework explains how moral autonomy can be fully appreciated when all pre-defined conditions are met. For example, when person X has insufficient accessibility to a digital device in the context of decision-making (given that all other conditions are met), the person is considered not autonomous. Now, when the same person does have sufficient access to a digital device, the supposition implies the person is morally autonomous (ceteris paribus). On that account, the supposition for all comparisons implies that every construct is positively (cor)related to HMA. However, we do not know the exact relationship in mathematical terms. The relationship can mathematically be defined as linear regression or a parabolic equation that is either convex or concave. Nevertheless, this is an enormous oversimplification of something normative. The aforementioned blue and orange scatterplots, representing perception and importance statements, respectively, are computed to analyse the presupposed correlation between HMA and all other constructs. The reason we merely analyse the relationships between HMA and all other constructs is multifaceted. Firstly, we generally analyse the HMA of end-users of BAIT in this study, and this is our main focus. Second, the other theoretical constructs are subordinate to HMA, as HMA is the overarching dimension. On that account, we study the relationships between HMA and all other constructs.

It is important to note for this subsection that the interpretations of the graphs entail a high uncertainty. We can identify irregular patterns to draw preliminary conclusions from this analysis. However, this does not imply a generalization or inference regarding the sample or population. Hence, we only identify patterns to demonstrate the potential ability the HMA survey possesses. In future studies with larger sample sizes, it can be used and analyzed with more certainty.

Relationships on perception statements

This subsection presents the graph of one exemplary relationship: critical questioning and HMA (see fig. 6.9). Based on the input of the respondents, the dots represent the mean score for the constructs. As a first exploratory step, we identify a positive relationship between both constructs. The dispersion is relative to other constructs on average. We are not able to derive conclusive evidence from this exploratory analysis.

A quick view of all other graphs shows that most of the regression lines are either positive or slightly negative. Those graphs are all included in section 8.3. As it was supposed, the graphs predominantly show a similar relationship as the theoretical constructs imply. In almost all cases, we observe a positive relationship between the constructs and HMA; however, some show clearer patterns than others. The graphs indicate that, in the context of BAIT, respondents substantially perceive the relation between the constructs similar to the assumed theory. Two graphs show a constant relation with HMA: accessibility and getting human high-fidelity expertise (see fig. 9 and fig. 11 in section 8.3). The graphs show a slightly negative regression line with high dispersion. In both graphs, we generally observe two clusters of data points with a similar distribution on the x-axis. This pattern may imply that the respondents perceive an equal amount of HMA on varying accessibility and getting human advice. This is is in contradiction with the underlying assumption of the theory. However, larger sample size may result in a positive relationship between both constructs, similarly to all other relationships.

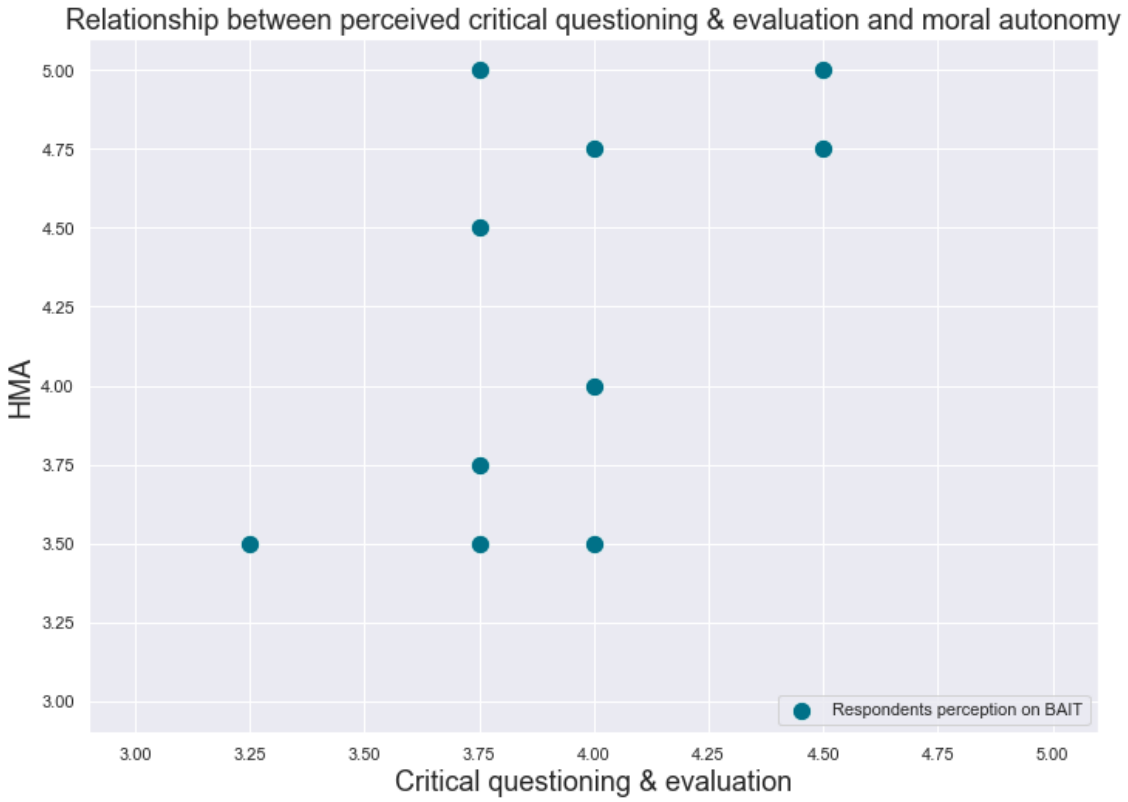


Figure 6.9: Perception of critical questioning and HMA

Relationships on importance statements

The relationship between the importance constructs reveals whether the respondents value the features of BAIT similarly to the theoretical framework's underlying assumptions. Hence, the statements on importance show whether respondents postulate similar relationships between constructs as hypothesized earlier. The theory behind the constructs implies that scoring similarly low on one of the constructs results in low HMA and vice versa. In fig. 6.10 the relationships between HMA and critical questioning and evaluation is visualized for the importance statement. Similarly to the perception statements, this graph indicates a positive relationship between both constructs. Therefore, we may conclude that - based on the preliminary findings of this study with a small sample size - respondents postulate a similar judgment on HMA and critical questioning equal to the underlying theory. This disclaimer is essential to all interpretations of the descriptive analysis.

A quick view of all other graphs shows that most regression lines are either positive or slightly negative (see section 8.3. We observe an equal number of positive relationships between constructs, with two graphs that demonstrate a constant (slightly negative) relationship (see fig. 19 and fig. 21 in section 8.3). Both present a slightly negative regression line with a similar dispersion. The continuous line starts high in both graphs, with a scattering of data over the HMA dimension. This pattern implies that both of the constructs score exceedingly high, with varying importance of HMA. The accessibility, tractability and critical questioning graphs (See fig. 19, fig. 20 and fig. 23, respectively) show a close relationship between the constructs and HMA. This similarity implies, for example, that respondents who score low on accessibility importance also score low on the HMA importance, and vice versa. Especially the critical questioning graph (fig. 6.10) demonstrates almost an identical line as the theoretical relationship. This implies indeed that people who value critical questioning importance high also appreciate their HMA high.

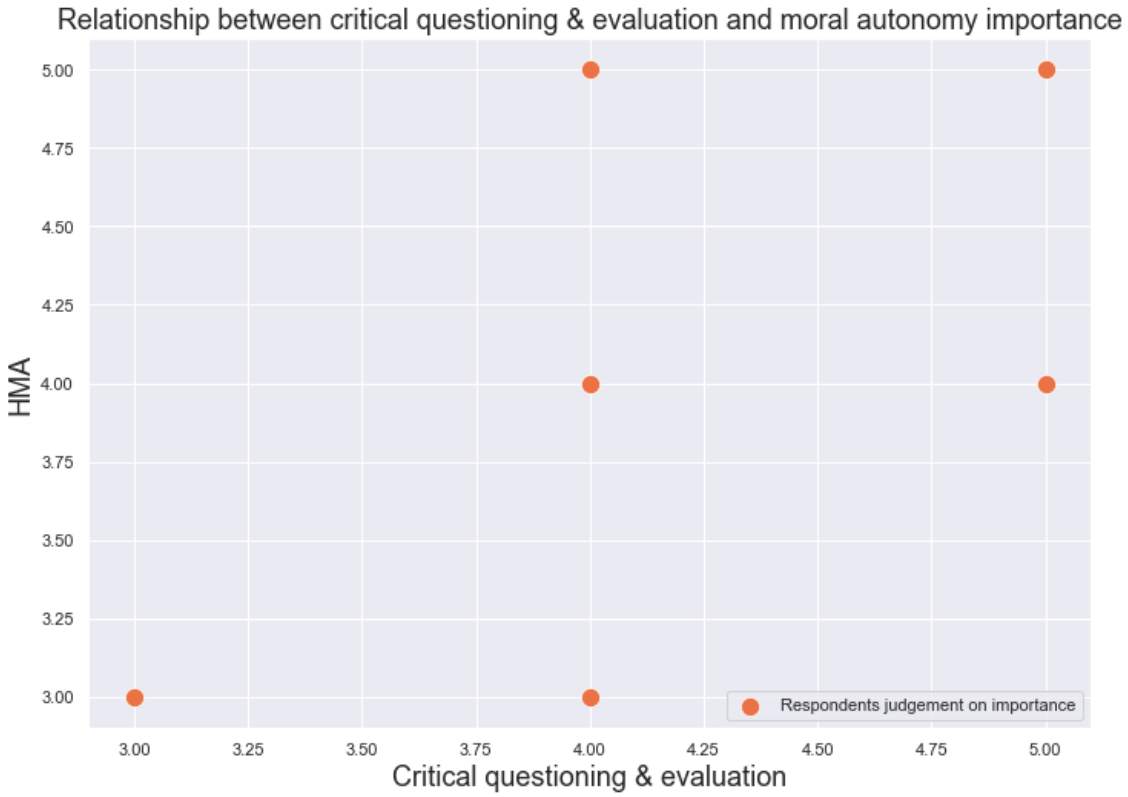


Figure 6.10: Importance of critical questioning and HMA

6.2 HMA guidelines to improve BAIT

This section entails the guidelines to improve BAIT in terms of HMA. We propose recommendations for each of the theoretical constructs we determined earlier in chapter 3. Thereby, we build further on the same structure of constructs we also studied in chapter 5 and the previous section. The guidelines are formulated based on best practices from the literature and combine this knowledge, wherever possible, with influences from the descriptive analysis. We aim to propose customized recommendations that may accurately improve the technology in terms of HMA. Hence, it primarily consists of best practices from the literature and considers valuable insights from the empirical part of this study.

6.3 Improving accessibility

Best practices for accessibility

Van den Hoven (1998) defines accessibility as: "Understanding the system as a matter of fact". In the present day, it fits the characterization of explainable AI partially. Explainable AI entails five main concepts of which only the former two show conceptual overlap with accessibility: transparency, causality, bias, fairness and safety (Wierzynski, 2018). Transparent AI is specified as an AI that can be understood and evaluated by humans (Goodman & Flaxman, 2017). This description implies that an AI should present all elements that play a factor in the prediction by the model. Most literature defines causality as understanding the underlying phenomena from explanations provided by the model based on the data (Hagras, 2018). However, researchers often put contemporary AI models within the narrative of performing low whenever they are explainable and vice versa, as illustrated in fig. 6.11. This description of AI models is, according to Rudin (2019), a myth. Especially when the model operates on structured data with clear representation and relevant features.

Explainability is, according to Shortliffe and Sepúlveda (2018), an indispensable feature of (clinical) DSS. The prerequisite substantiates that clinical experts need to understand the groundwork of advice provided by the system to manage trade-offs in triage situations. Even more, Montani and Striani emphasises the fact that there is a high need for explainable DSS, regardless of the AI

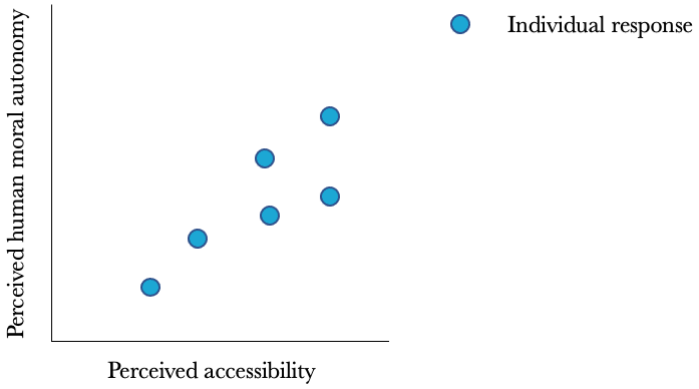


Figure 6.11: Fictitious performance-explainability trade-off (Gunning, 2016)

on which it runs. We can conclude from this that many domains do not accept black boxes unconditionally. Hence, explainability has been a central theme recently. Additionally, the demand for explainable models is increasingly growing. To improve accessibility by proxy of higher explainability, the framework of Gunning (2016) will be used to pinpoint features that can be improved (see fig. 6.12). The framework explains the interaction of BAIT, subdivided into two entities: the model and the interface. Furthermore, it illuminates how the system interacts with the expert, who eventually makes a decision.

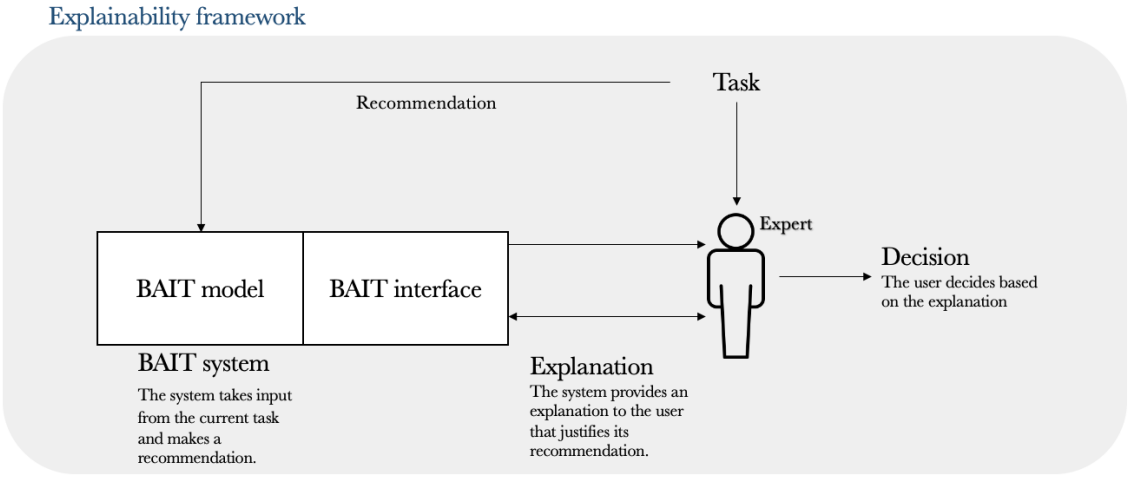


Figure 6.12: Explainability framework
Derived from (Gunning, 2016)

The accessibility of a DSS is dependent on the integration of inherently explainable models. BAIT utilizes a relatively transparent tool, namely discrete choice analysis. This forms the statistical foundation to design the model and interface. To formulate the recommendations to increase the explainability of BAIT, the principles of Kulesza, Burnett, Wong, and Stumpf (2015) will be synthesized with our understanding of BAIT. By doing so, we aim to provide appropriate and applicable points of improvement for BAIT to respect HMA on this specific sub-domain.

Firstly, Kulesza, Stumpf, Burnett, and Kwan (2012) emphasises the importance of mental models, which fulfil a supportive function for humans to understand the system better as a matter of fact. Mental models cover relationships between concepts and related systems. There are two categories of mental models: (1) functional models that show a basic understanding of the end-user and structural models that imply a deeper understanding (Norman, 1983). Building mental models require transparent systems and comprehensible instructions (Rogers, Sharp, & Preece, 2011). This seems to be feasible with the application of BAIT. In the case of BAIT, the advice to the end-user could be visually sub-divided into various components, in which the system shows

the connecting relationships. This structural way of providing knowledge on how the system works appears to be highly beneficial when it comes to computer self-efficacy, and a general understanding of end-users on the system (Kulesza et al., 2012). This methodology is especially valuable when end-users are encouraged to make mental models themselves based on the provided information of BAIT.

Secondly, soundness determines the quality of explanations extensively. Soundness is referred to as "the extent to which each component of an explanation's content is truthful in describing the underlying system" (Kulesza et al., 2013). This definition implies the system should not be presented as less complex. BAIT may accomplish this in terms of the aforementioned mental model. For each step within the model, the soundness of the explanation can be differentiated with the inherent mathematical computations. Consequently, one can evaluate how well the descriptions fit the essence of the operations conducted by the system. The explanations should thereby reflect the internal process to define it as a sound explanation.

The perceived trade-off made with soundness is the third principle, completeness. Completeness involves "the extent to which all of the underlying systems is described by the explanation" (Kulesza et al., 2013). Supporting intelligibility by providing explanations will increase the acceptance rate of the target group and, additionally, contribute to their accessibility (Lim & Dey, 2009). To design precise mental models, one should bear in mind to provide complete information on the system infrastructure. Only then it becomes feasible for non-specialists to understand the relationships between different parts of the system. This can be intercepted by focussing on the intelligibility of information. For most systems, it remains a challenge to reveal the extent of data that needs to be presented. This can be explored by asking appropriate questions to end-users. Moreover, intelligibility questions may serve to reach completeness of explanations to end-users: why (not), how, what if, what else, etc. Lim and Dey (2009) provides more recommendations to support intelligibility by providing explanations in various ways. To conclude, it is paramount to understand how complete information may contribute to the system's intelligibility.

Interpretation of descriptive analysis

The descriptive analysis found that respondents scored accessibility of BAIT 3.9 on average but found it relatively more important: 4.4. The mean difference between the perception and importance average respectively appeared to be significant (see section 8.3). Although the mean perception is relatively high with four out of five points, there is still room for improvement. The first three statements of this construct mention the degree of accessibility to relevant information, understanding how the system works, and presenting correct information. Those three scored equally according to the sample. If we analyze those statements more profoundly, we notice the connection with the proposed recommendations. By making mental models of the system and providing complete and sound information, we aim to improve end-user's understanding and provide relevant and correct information. On that account, the technology can be improved and measured accordingly. Thereby it is the perceptions can be measured appropriately, as they are conceptually aligned.

6.4 Improving Tractability

Best practices for tractability

In a formerly mentioned paper on explainable AI, Kulesza et al. (2015) already explained end-users must be able to "drill down" through the system to understand the technology more thoroughly. Van den Hoven (1998) defines tractability as "the ability to keep track of information within the system". Hence, end-users ought to be able to trace the advice of BAIT to more profound substantiations. This state is only possible on the premise, as mentioned earlier, of accessibility. End-users are only able to keep track of information when the accessibility of the technology provides sufficient information. This subsection explains how we can improve BAIT in terms of tractability.

Additionally, tractability requires the technology to be transparent (Dattachaudhuri, Biswas, Sarkar, & Boruah, 2020). For machine learning, transparency is generally defined as sharing the code that is used to train the model and determine the parameters (Haibe-Kains et al., 2020).

For BAIT specifically, this is distinctly different, as it is fundamentally not similar to machine learning. Nevertheless, it is paramount to emphasise the importance of transparency to enable high tractability of BAIT. Therefore, the criteria of Tan, Ng, and Quek (2008) will be used to define fitting criteria for BAIT to improve its tractability.

Firstly, to provide a piece of tractable advice to end-users, it is paramount to present it with human-like reasoning (Tan et al., 2008). Human-like reasoning refers to the demonstration of the advice using precise specifications the way humans generally perform the think-process (Fahlman, 2011). BAIT can fulfil this criterion by presenting appropriate high-level overviews of how the usual decision-making is performed to the end-users. Additionally, the presentation of human-like reasoning should resemble the familiar procedure experts use to make decisions. For each consecutive variable that contributes to the model's advice, it can be shown how it contributes to the direction with clear descriptions. Providing a high-level overview with human-like reasoning makes it easier for the end-user to identify the patterns, as the system follows a similar think-process.

To build further on the premise of human-like reasoning, the model should present the advice step by step (Tan et al., 2008). This recommendation can be related to the mental models as discussed in section 6.3. By providing step-by-step advice in which the end-user is encouraged to follow the line of reasoning by the model, they will be able to understand how the direction originates. To conclude, a step by step inference supports the understanding of the system by the end-user more thoroughly.

The last recommendation to enhance tractability is the usage of familiar terms and justifications of the domain the end-user works in (Tan et al., 2008). To ensure tractability is applied correctly, BAIT should define the variables in reasonable terms. Moreover, explanations on each variable should be given and justified in an appropriate language. Only then users can track down the advice given by the DSS. To conclude, tractability can be enabled technically, but intelligibility remains a challenging task.

Interpretation of descriptive analysis

The descriptive analysis revealed that the mean of the perception and importance statements appeared to be 3.77 and 4, respectively. This difference is relatively small and the t-test proved the difference between the mean of both measurements is insignificant (see section 8.3). However, with a bigger sample size, we hypothesize the difference can be measured significantly. Albeit not being significant, the perception statements score lower than the importance statement. On that account, we take a look at how the recommendations may improve the technology. A more profound observation of the heatmap (see fig. 6.5) shows that TRAC4 scores relatively the lowest among the statements. This statement mentions the ability to trace information within BAIT. We assume that the recommendations (human-like reasoning, step by step inference and usage of familiar terms) may substantially improve the tractability of BAIT if applied correctly.

6.5 Improving get human high-fidelity expertise

Best practices for getting human expertise

Getting high-fidelity human expertise is subordinate to time pressure. Van den Hoven (1998) defines this theoretical construct as being able to get out of the artificial epistemic niche and consult peers before making decisions. When the end-user is not able to get out of the so-called artificial epistemic niche, Van den Hoven (1998) argues that the technology can be classified as a 'narrowly-embedded system' (i.e. the end-user is excessively dependent on the features provided by the technology). From this, we can conclude that the end-user must always get out of the artificial epistemic niche and reach out to colleagues before making a decision. This construct is generally put in the context of time limitation in decision-making, as expressed by Price, Walker, and Wiley (2018) in a military context:

"*Speed has always constrained commanders' decision time and that trend is accelerating. This will generate huge pressure to rapidly understand and act upon the information. AI holds the promise to help with this problem and creates a temptation to value the risk posed by the enemy*

much higher than the risk posed by algorithmic error and opacity. This could lead to the operational deployment of prototype AI systems that have not undergone rigorous evaluation and testing".[1]

In chapter 3 it was already explained how, in the context of supported moral decision-making (TDP2), the human agent suffers from cognitive underload and is expected to decide in time (Waa & Diggelen, 2020). Hence, time fulfils a significant role to enhance the ability to get advice from peers.

Getting high-fidelity human expertise is highly context-dependent. Some situations reveal more explicit moral dimensions than other situations, but both choice situations raise choice alternatives that can be classified as 'right' or 'wrong' (or something in between) (Chorus, 2015). Hence, moral decision-making requires identifying the moral dimensions of the decision-making. However, it is more feasible for BAIT to classify the decision-making based on complexity for practical reasons. Hughes and Young (1990) shows with their results in clinical decision-making how agreement among clinical workers diminishes as task complexity increases. Hence, every decision-making should be clustered based on complexity to identify the most doubtful cases. This can principally be done in two ways. First, BAIT could serve the role as pre-selection in decision-making. The treated cases are clustered based on complexity levels as determined by the model. The most complex issues would then require the end-user to discuss it with peers. Secondly, cases can be clustered throughout the decision-making by using thresholds for the advice percentage. If a case exceeds the maximum or minimum threshold, the system recommends the human discuss it with peers. These recommendations may increase safety in the decision-making. By doing so, we ensure the end-user has enough time to get high-fidelity human expertise.

BAIT should merely remain as a supportive tool by sticking to its supportive function. Parasuraman et al. (2000) exemplified with their table on levels of automation how machines can operate on various automation levels (see fig. 2.7). If the aim is to enhance the HMA of end-users, it is paramount to not let the DSS act as an autonomous (artificial) agent in the decision-making. In contrast, it should rather act as an automation tool (O, McNeese, University, Carolina, & Barron, 2020). DSSs that deal with decision-making in dynamic environments with a degree of uncertainty should not advance to higher levels of automation (N. B. Sarter & Schroeder, 2001). Hence, the integration purpose of BAIT is paramount to prevent abuse by a misconception. It is paramount to ensure BAIT does not exceed certain automation levels to enable the option of getting human expertise.

The theoretical construct of getting high-fidelity human expertise is a self-contained criterion. BAIT should respect the HMA of the end-user should by enabling the option to get advice from peers if requested. Kant argues moral autonomy is a combination of freedom and responsibility (Wolff, 1998). This argument implies that time is part of freedom to act upon one's principles truthfully, as many philosophers argue (see chapter 3). The ability to get advice from peers is only possible when the aforementioned moral dimensions are clear to the end-user and when there is sufficient time to fulfil the decision-making (N. B. Sarter & Schroeder, 2001). This pre-requisite includes the ability to get support from peers whenever it is deemed necessary. To conclude, the last recommendation of this theoretical construct is to ensure the DSS guarantees the end-user the ability to get advice from peers.

Interpretation of descriptive analysis

In the descriptive analysis, the mean for perception and importance statements appeared to be 3.72 and 4.82, respectively. This difference is relatively high to other constructs and appeared to be highly significant according to the t-test (see section 8.3). The statements within this construct are mainly concerned with encouraging to contact others and the interchange of ideas. The formulated recommendations cover the classification of cases based on complexity, not exceeding automation levels and preserving sufficient time. Therefore, we hypothesise that respondents will score higher on the statements as they get more time and bear good task responsibility. Hence, the recommendations may substantially improve the technology in terms of this construct, which is statistically measurable.

[1] This is stated in the context of military warfare

6.6 Improving identify & empower human expertise

Best practices for identifying & empowering human expertise

The fourth theoretical construct entails identifying and empowering human expertise among end-users with the application of BAIT. Van den Hoven (1998) argues individual autonomy and responsibility deteriorates when available data is presented to end-users who are less experienced. This decay can especially be problematic when the system is built on the knowledge of human experts during the construction of the model and used by less experienced end-users. DSSs are primarily designed to present knowledge to enable the end-user to make reasonable decisions (Alexander, 2006). Hence, it is paramount to emphasise human expertise using BAIT to (1) truly fulfil its function as a supportive tool and (2) embody the HMA of end-users.

The usage of digital systems undoubtedly affects human expertise, in which automation bias fulfils a significant role. When systems provide consistently accurate recommendations, human experts tend to rely on the system increasingly (Cummings, 2006). When users consider the proposals by BAIT to be a deterministic unit of information for the decision-making, they likely disconnect from the traditional think-process. Dearden (1972) referred to this latter concept in terms of moral autonomy as: "A person is "autonomous" to the degree that what he thinks and does cannot be explained without reference to his own activity of mind".[1] Ensuring the knowledge of end-users is sufficient can be enabled by avoiding over-reliance by humans on BAIT. BAIT can serve multiple purposes (e.g. providing insights on decision-making, automation, etc.) with varying consequences. It is therefore essential to acknowledge the differences in usage by, for example for juniors and seniors. Secondly, it is crucial to recognize the end-user must always be able to execute the decision-making independently.

A digital supportive tool focuses on the development of knowledge representation by a computer to support and advise humans (Salisbury, 2019). This function also remains an objective for BAIT. Anderson, Bloom, et al. (2001) set the ground for illustrating types of knowledge by differentiating between them (based on taxonomy of Bloom et al. (1956)). This is shown in fig. 6.13. BAIT does already provide factual knowledge by providing the weights for criteria and the relative importance among them. Additionally, it could provide valuable insights by focusing on conceptual knowledge too. Conceptual knowledge entails both implicit and explicit understanding of ruling principles and relations between parts of information within a domain (Rittle-Johnson & Alibali, 1999). In the study of van der Waa et al. (2021) the distinction is made between two different types of explanations. Those are compared employing an experiment: rule-based and example-based explanations. This study is exemplary to identify the best types of explanations by a system to increase the end-user's understanding. Therefore, to provide more profound insights on the decision-making, BAIT may show relations between different variables (e.g. correlations) and display how choices of the DSS develop throughout a period (including a comparison with the actual choices made by the end-user). To conclude, differentiating between types of knowledge empowers human expertise.

Interpretation of descriptive analysis

In the descriptive analysis, we found out the mean for perception and importance statements are 3.5 and 4.45, respectively. Utilizing a t-test, the difference between the values appeared to be significant (see section 8.3). This difference is relative to other constructs high. The statements within this construct generally mention the competency of end-users, identification of experts and improvement of knowledge. The difference between the means of both measurements indicates the respondents usually are concerned with human expertise in the context of BAIT. Hence, by testing how knowledge can at best be presented, we aim to improve the so-called empowering of human expertise. Moreover, identifying different types of knowledge can be used to provide advanced knowledge to the end-user. Overall, this will hypothetically lead to enhanced human expertise.

[1] This is stated in the context of learning modules in education. Complex decision-making is, however, also concerned with learning processes

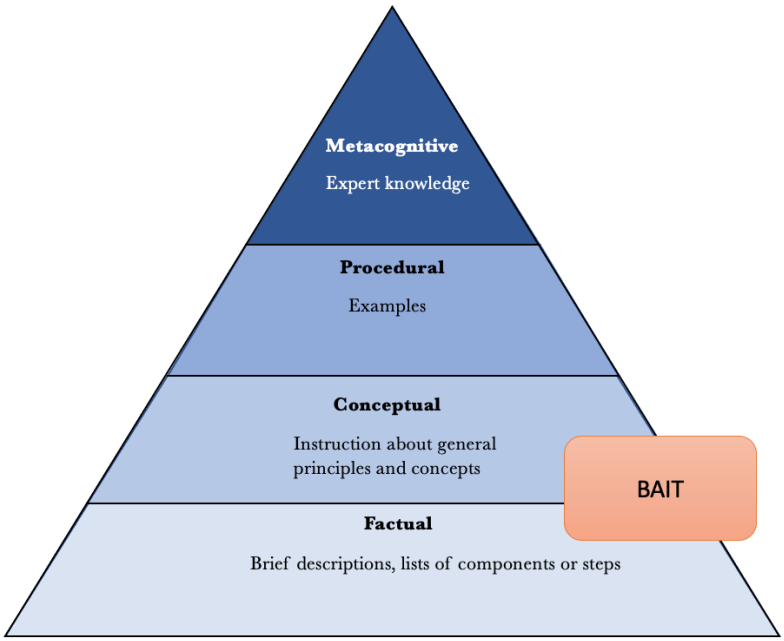


Figure 6.13: Types of human knowledge
Derived from (Anderson et al., 2001)

6.7 Improving critical questioning and evaluation

Best practices for identifying & empowering human expertise

Participating in the last condition of moral responsibility, Van den Hoven (1998) explains that, given the inability of identifying and empowering human expertise, the room for critical questioning and evaluation gets severely restricted. The end-user can be motivated to think and evaluate the decision-making based on two premises critically. Firstly, the system should present sufficient information intelligibly. Second, human expertise should always remain a top priority. System understanding is referred to as the ability to know what the system consists of and how it behaves internally (van der Waa et al., 2021). This theoretical construct is subdivided into two recommendations: the end-user must be informed on different types of decision-making to evaluate the system critically and use validation methods to examine the system.

Critical questioning of the system coincides with a high system understanding by the end-user. Algorithmic decision-making is founded on statistical models (e.g. BAIT is based on discrete choice analysis) to provide insights on decisions more objectively to improve the decision-making (Chen, Chiang, & Storey, 2012). However, Shollo and Kautz (2010) demonstrates how end-users, who routinely use algorithmic decision-making, tend to lose the sense of the information processing that traditionally forms the basis for knowledge. Hence, it is vital to provide information to end-users on how algorithmic decisions and human decisions differ from each other and how they can be used interchangeably to evaluate decision-making critically. BAIT may enhance critical evaluation by informing end-users on the different types of decision-making, as illustrated in fig. 6.14. Both kinds of decision-making hold their limitations but can strengthen each other. To conclude, human decision-making characteristics should improve the end-user's capability to critically evaluate the technology.

Being able to question the system critically, and hence, algorithmic decision-making can generally be done with validation methods. Borenstein (1998) provides four practical tests to validate the DSS by end-users: (1) face validation, (2) predictive validation and (3) User assessment. The purpose of face validation is to realize a product that is consistent with the user's view of the problem (O'Leary, Goul, Moffitt, & Radwan, 1990). It is an early response for developers to evaluate whether the system meets the requirements as defined by the end-users. Predictive validation entails using test cases in which the results are known to compare the results between both. The last method is the user assessment. This method requires an evaluation by end-users, in which the

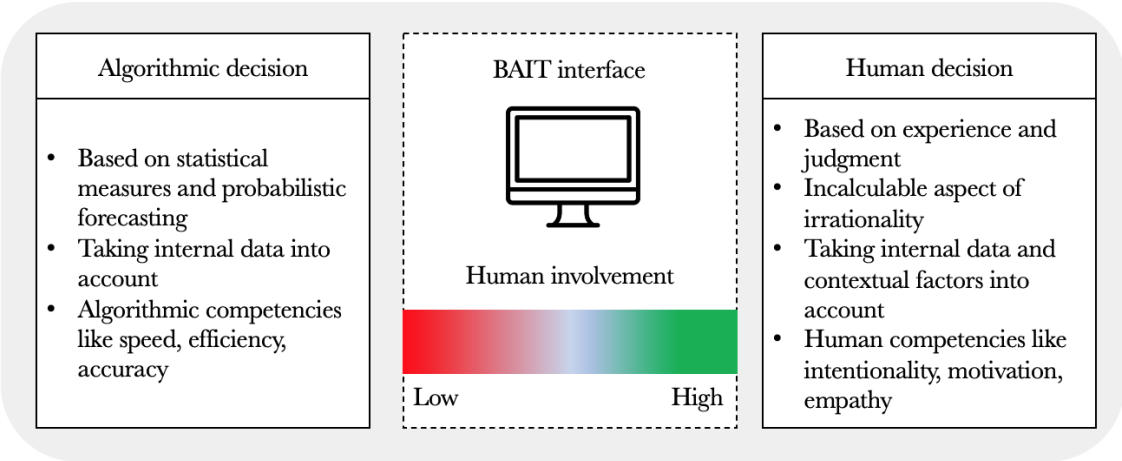


Figure 6.14: Human and algorithmic decision-making
Derived from (Bader & Kaiser, 2019)

aim is to determine whether the DSS can be used in decision-making (Gass, 1983). The objective of this method is to understand the applicability and evaluate the impact of the system's characteristics (Borenstein, 1998). The methods, as mentioned earlier, are illustrated in fig. 6.15 and show in what order they (iteratively) can be conducted. To conclude, validation methodologies support both developers and end-users in critically assessing the technology and its impacts on decision-making.

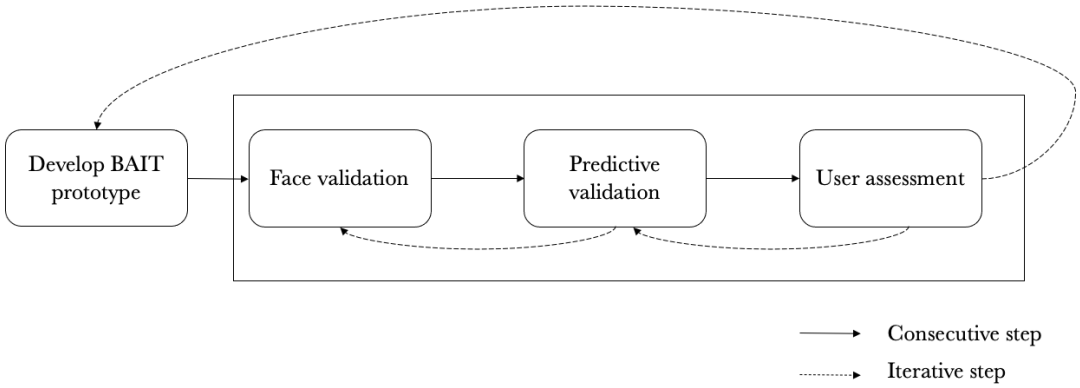


Figure 6.15: Potential validation steps for BAIT
Derived from (Borenstein, 1998)

Interpretation of descriptive analysis

Based on the input of the small sample size, the mean score for perception and importance statements appeared to be 3.9 and 4.27, respectively. This difference is relatively small and, unsurprisingly, appears insignificant according to the t-test (see section 8.3). Albeit the perception of BAIT appears to align with the importance respondents ascribe, the technology can be further improved. The problem owner can use the recommendations to inform the end-users on the different types of decision-making and ensure the model is critically built, based on integral cooperation with end-users with multiple validation steps. By doing so, we aim to retrieve higher perception scores for critical questioning and evaluation. To conclude, although the current mean difference is relatively low and insignificant, it can be improved to enhance the HMA of end-users.

6.8 Guideline matrix

This section presents the guidelines in a table form in which the results of both aforementioned sections are combined in an overview. The table is presented hereafter and consists of the theo-

Domain	**Recommendation 1**	**Recommendation 2**	**Recommendation 3**	**Practicality**	**Results**		
					Mean perception	Mean interest	Discrepancy
Accessibility	Encourage design of mental models by end-users	Create sound explanations for each component of the model	Provide complete information by means of varying explanations	4	3,9	4,36	0,46 (p-value = 0.039)
Tractability	Enhance human-like reasoning in information presentation	Present step-by-step computation of model	Use familiair terms and justifications of the end-users' domain	4	3,77	4	0,23 (p-value = 0.367)
Getting human high-fidelity expertise	Let BAIT perform preselection of decision-making and use explicit thresholds	Avoid higher levels of automation	Ensure sufficient time for end-user to get advice by peers	3	3,72	4,81	1,09 (p-value = $1.715*10^{-6}$)
Identify & empower expertise	Consider differences in expert level and ensure independence of end-user	Identify different types of knowledge to present		2	3,5	4,45	0,95 (p-value = 0.0004)
Critical questioning and evaluation	Inform end-users on different characteristics of decision-making	Use multiple validation methods to ensure critical assessment by both developers and end-users		3	3,93	4,27	0,34 (p-value = 0.204)

Table 6.1: Guidelines to improve HMA of end-users BAIT

retical constructs, recommendations, conceived practicality and main results from the descriptive analysis. Altogether, this can be used for further improvement of the technology.

Summary box chapter 7

The goal of chapter 7 is to demonstrate the potential ability of the measurement instrument and formulate points of improvement for BAIT. By conducting the descriptive analysis, we exhibit the valuable information the HMA survey provides. This enables researchers to find the relation between certain constructs and statements. The second part entails the guidelines to improve BAIT. The guidelines are primarily formulated based on best practices from the literature. Additionally, we show how the descriptive results connect to each of the recommendations. We demonstrate how the results from the empirical study contribute to improving the technology in terms of HMA. The following points summarise this chapter:

- The number of respondents was relatively low to form conclusive evidence
- We were able to measure the scores per statement and construct. This enabled us to compare them and identify interesting findings.
- The distribution of scores per construct are compared within constructs. Utilizing a t-test for each mean difference between the perception and importance statements. This enabled us to compute the significance of each difference.
- The scatterplots are valuable to find relationships between the hypothesized constructs. However, due to a low number of respondents, there is high uncertainty in the interpretation.
- The guidelines contain two to three recommendations per construct. Each recommendation is substantiated with notable scientific papers. Moreover, we included the understanding of the descriptive analysis to state the urgency of each improvement.

Chapter 7

Discussion

This chapter reflects the results of all sub-parts within this study, clarifies the scientific and practical contributions and mentions the limitations. In section 7.1 each part of this study is thoroughly reflected on by taking the methodology under a magnifying glass. Moreover, answers on the sub-questions are formulated. In section 7.2 we evaluate the interplay between social acceptance and ethical acceptability based on the results. In section 7.3 and section 7.4 we clarify the value of this study by explaining the scientific and practical contributions.

7.1 Reflection on study

This section entails the reflection on each part of this study and formulates answers on each of the sub-questions. By doing so, we aim to reflect on the decisions to understand the implications. Thereafter, we reflect on the content by answering the sub-questions. Altogether, this provides a coherent discussion of this study. The thematic structure of the reflection supports to mention each sub-question thoroughly. The visual presentation is shown in fig. 7.1.

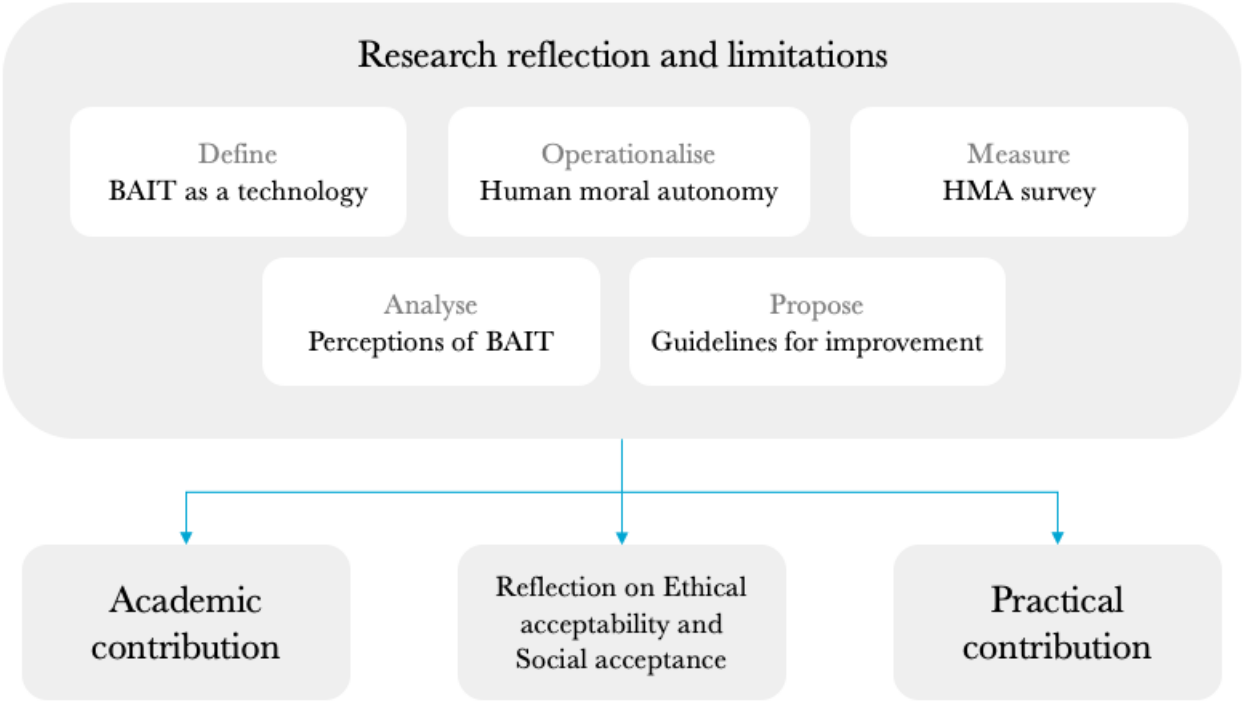


Figure 7.1: Discussion structure

7.1.1 Defining the technology

The first part of this study entails the definition of the technology in technical terms primarily. We discover the usage of BAIT in realistic settings and what implications this may have for the task allocation. This subsection aims to answer the following sub-question:

Research question 1

In what ways does BAIT support human decision-making in organizations and how does that affect task allocations between the human and the DSS?

The purpose of this sub-question is to understand BAIT as a technology, what it consists of and how end-users can use it in a decision-making context. The system provides advice to the end-user regarding the percentage of experts that would decide to follow up on the recommendation. The percentage refers to the share of respondents that agrees on a specific case. Hence, in contrast to most non-knowledge based models like machine learning, it does not define the advice in terms of performance or accuracy (Rudin, 2019). It merely refers to the behaviour of a specific group of respondents. Moreover, it cannot provide any information on cases in which the attribute levels are exceeded; the model does not support extrapolation. BAIT aims to help humans in decision-making by giving intelligible information on how end-user peers would likely behave in the situation. With this, BAIT distances itself from non-knowledge based systems that find elusive patterns in enormous data sets. It also distances itself from conventional expert systems that can only infer knowledge based on rigorous rules. To conclude, BAIT supports experts in their decision-making by solely providing advice based on a technique that extracts information from hypothetical trade-offs.

The allocation of tasks depends on the hypothesised level of automation, as this points out the extent to which it is assumed that humans will rely on technology. In fig. 2.7 we illustrated how an increased level of automation allocates more tasks and responsibilities to the artificial agent. BAIT can suggest one alternative (level 4), but it will likely act between levels five and seven with further developments. According to Parasuraman et al. (2000), this range indicates a machine to be a partially autonomous agent. With this information, we located BAIT on the second team design pattern (TDP2) of Waa and Diggelen (2020): supported moral decision-making. To put it more concretely, the artificial agent performs the task by fulfilling its role as a decision-maker (e.g. grant a subsidy, participate on a public tender, etc.). The human agent controls the variables and evaluates the decision of the model. To conclude, the task allocation implies the human agent must supervise the model properly to ensure moral awareness and have a sufficient understanding of the ethical implications. The artificial agent must provide adequate time and explanations to enable the human agent to take over the decision-making.

We characterised BAIT based on its current technical features and the available literature. However, BAIT has not been implemented in the decision-making of any organisation yet. Hence, we could only understand how BAIT may support organisations' decision-making and task allocation by comparing it with the available literature. On that account, we used a broad range of automation levels (four to seven) to specify the technology. However, Further development of BAIT may induce different exhibitions of the system.

Limitations and implications

As BAIT is still a novel technology, many features have not been studied yet, although it shows similarities with other technologies. In section 1.1 we illustrated how this study, within the framework of Floridi et al. (2018), only touches upon one sub-domain: autonomy. All other sub-domains within this framework have been excluded from this study. This decision puts forward the first main implication. Nevertheless, other sub-domains do play a varying role within this study, implicitly or explicitly. Moreover, as BAIT still is in the development phase, we could only characterise the technological features in conceptual terms. Further studies may characterise BAIT more specifically by examining its applications.

7.1.2 Operationalising HMA

The second part of this study entails the operationalisation of HMA as a concept. By operationalising the normative nature into measurable concepts, the aim is to put the conditions of HMA more concretely. This subsection answers the following sub-question:

Research question 2

What is a philosophical definition of human moral autonomy in the context of a DSS like BAIT?

The characterisation of HMA is a crucial step, as many philosophers have initiated their understanding of moral autonomy (Dworkin, 2015). We decided to use the concept of human moral autonomy (HMA), as it merely refers to the moral autonomy of human agents. The definition of HMA is eventually defined based on two complementary papers by Dworkin (1981) and Sternberg (2012). The synthesis of both concepts led to the following definition: "A person is morally autonomous if and only if he bears the responsibility and authority for moral supervision on the situation and the decision support system". This definition is our understanding of HMA in the context of BAIT. Accordingly, we formulated conditions for the definition of HMA to realise the operationalisation.

It is paramount to include relevant and suitable conditions to build on the definition of HMA. Hence, we used the paper of Van den Hoven (1998) to set up the theoretical framework. The main conditions in computerised environments are (1) inscrutinizability, (2) pressure condition, (3) error condition and (4) critical questioning & evaluation. We assumed those constructs to be determinants for the moral responsibility of the end-user. The conditions contain sub-conditions that can be made measurable. Hence, the goal of this research is fulfilled by defining the conditions and sub-conditions of HMA. Altogether, we were able to formulate a philosophical definition of HMA and operationalise this normative concept. However, we do not claim this framework to contain the solely legitimate selection of conditions that determine the autonomy of decision-makers. We encourage other researchers to define different conditions to measure HMA. By doing so, researchers can compare varying frameworks to extract the most appropriate variables for empirical studies.

Limitations and implications

The framework is eventually constructed based on the paper by Van den Hoven (1998). Hence, we only restrict ourselves to the theoretical constructs as defined in this paper. This is illustrated in fig. 3.6. The selection of constructs has undeniable consequences for this research. We limited ourselves to a number of conditions to operationalise HMA. Therefore, an important disclaimer for this part is the theoretical limit of the concept of HMA. To conclude, the theoretical framework on HMA is constructed based on just a couple of papers. The main purpose of this activity is to maintain conceptual consistency and understand its limitations for broader interpretations. However, this selection obliged us to limit the operationalisation to a sub-part of HMA.

7.1.3 Measuring HMA

The third part of this study entails the measurement instrument of HMA in the form of a survey. This directly builds further on the previous operationalisation of HMA and aims to quantify the constructs to measure them. This subsection answers the following sub-question:

Research question 3

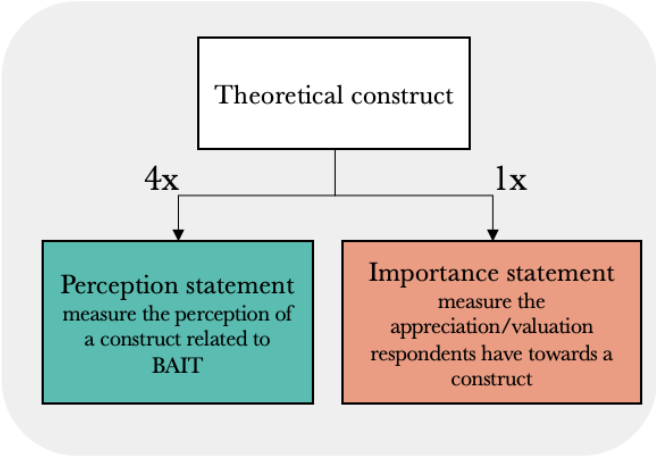
How can we measure the degree to which DSSs like BAIT respect human moral autonomy using a survey?

As it is not feasible to take into account all the constructs of the HMA framework (see fig. 3.6) into the empirical study, we decided to take only six constructs: (1) HMA, (2) accessibility, (3)

tractability, (4) getting high-fidelity human expertise, (5) identify & empower human expertise and (6) critical questioning and evaluation. HMA is the overarching construct, the latter five are subordinate to HMA. The constructs are deemed most practical and applicable in the context of BAIT. Moreover, they fit appropriately into the aforementioned problem description. Potential users indicated during the pilot surveys (section 8.3) they worry about the development of expertise (construct no. 4,5,6), question the system characteristics (construct no. 2 and 3) and question their responsibility as a whole. To conclude, the constructs are not only deemed important by the literature, but also by potential end-users.

Measuring the degree of HMA is done by extracting the perceptions of decision-makers. Consequently, we built the HMA survey based on the premise of factor analysis. Hence, each construct includes four statements to measure the perceptions of end-users on BAIT. Additionally, for each construct, an importance statement is formulated to measure the degree of urgency. On that account, we enable ourselves to compare perceptions in the context of BAIT with the importance respondents ascribe to each construct. Any discrepancy between perceptions and importance is used to measure the degree to which BAIT respects HMA.

Figure 7.2: Types of statements



After completing the initial version, we conducted pilot surveys to pretest the survey. We did this for two purposes. Firstly, we aimed to pre-validate the questionnaire to test whether it is a suitable tool. Secondly, we aimed to obtain additional information on the opinions of potential respondents. The second purpose appeared helpful, as it provided further insights into what end-users care about in digital systems. The pretest is done by conducting a semi-structured interview in which respondents are questioned on the constructs. The presentation of the pilot survey is included in section 8.3. By doing so, we were able to test respondents' knowledge on similar constructs and qualitatively comprehend their understanding of HMA in the context of BAIT.

Although more respondents are deemed necessary to validate a survey significantly, we still decided to use factor analysis. The number of respondents appeared to be too low to validate the correlation structure of the data. We proceeded with the factor analysis to validate the survey. By continuing the validation, we were able to extract new components mainly represented by differing constructs. Only one hypothesised construct appeared within one component, containing two statements: the HMA construct. With all other components, we were obliged to combine constructs. Although the final components do not reflect our hypotheses appropriately, they measure the critical sub-conditions we argued in the theoretical framework. We suggest future studies reach a bigger sample size with similar respondent characteristics. This is crucial to ensure the validity of the HMA survey.

The final components do not distinctively represent the sub-conditions as we hypothesised within our HMA framework and survey. We notify all of the accessibility, critical questioning and HMA statements are included within separable components. Hence, based on the current input of the respondents, those constructs are increasingly coherent with the new component structure. The final form of the HMA survey predominantly measures the perceived critical questioning,

HMA and accessibility of respondents within a technology. Tractability, identify expertise, and get human expertise appear with just one statement in the final component structure. Albeit those constructs fulfil an inferior role, we decided to include them to suitably reflect our initial selection. However, we may argue that the former three constructs are strongly correlated and achieve higher validity. Hypothetically, future research may elaborate on these three constructs. However, this sub-selection only contains two sub-conditions, as HMA is the overarching construct. We dissuade researchers from elaborating on this number of sub-conditions, as it is too little to measure perceived HMA entirely.

Limitations and implications

The factor analysis did not result in any significance due to the limited number of respondents that we could reach. We earlier explained that finding an appropriate number of respondents remained to be a challenge. The potential number of people we tried to get is 45, by which 11 people have responded. We found out that most people at these organisations indicated they are extremely busy with their regular job (especially the surgeons and intensivists). Others pointed they feel fed up with all surveys they have to fill in, especially since the covid-19 pandemic. Although we did not reach the desired number of respondents, we could still extract exciting findings from the data.

We were able to validate the survey utilising factor analysis. However, this validation contains some snags. Firstly, we could not validate the correlation matrix due to a low number of respondents. Second, we combined statements from various constructs to form new components. The final components do not perfectly reflect our hypothesised constructs from the HMA framework. This has implications for the interpretability of the survey results. Elaborating on this study requires understanding all of the underlying constructs as we defined within this study. On that account, it is paramount to keep the underlying fundamental theories of the HMA survey transparent. We propose for future studies to distinguish the hypothesised constructs similarly, improve the statements based on the findings with the results from this study and reach a bigger sample size.

7.1.4 Analyse HMA perceptions of BAIT

The fourth part of this study entails the descriptive analysis of the given input by potential end-users. Hence, this forms a demonstration of the HMA survey and the potential applications it has. The aim is to analyse perceptions of HMA in BAIT and identify discrepancies between perception and importance. This subsection answers the following sub-question:

Research question 4

4. To what extent does BAIT respect the conditions of HMA according to end users?
   (a) What importance do end users ascribe to the philosophical conditions of HMA?
   (b) Does BAIT respect the philosophical conditions of HMA in the perception of end users?

The descriptive analysis resulted in four sub-parts: sample information, high-level analysis, distribution of scores and the relation between HMA and all other constructs. Within the high-level analysis, we aim to identify patterns between statements, constructs and respondents. We observe with a quick overview on the heatmap that the statements on HMA (MA1,2,3,4) score relatively high. Moreover, we notice that some statements on getting high-fidelity human expertise and identifying & empowering human expertise score lower. Especially EXP2 ("Using the Councyl model makes it more difficult to recognise (new) experts in my department") scores lowest among all of the statements. The interviewees of Deloitte were concerned with the development of human expertise when using models such as BAIT; this statement relates to what they referred to as 'industry sensitivity'. We also computed correlation matrices within constructs (see fig. 6.6) to determine the coherence. The statements on HMA contain higher positive correlations (approximately 0.30), whereas, for accessibility, we observe negative correlations among the constructs. Overall, the correlations are not sufficient to formulate any conclusions based on the preliminary findings. In section 8.3 we also included a dendrogram based on the heatmap, which we clustered

based on the organisation. We identified a relationship among the respondents based on the organisations, especially the respondents of UMCG seem to stick together. However, more respondents are required to formulate more reliable conclusions on this.

The third sub-part of the descriptive analysis consists of the distribution of scores. Section 8.3 includes all the distribution scores for each of the constructs. In chapter 5 we present the results of the overall scores on the perception statements and importance statements. The perception statements have an average value of 3.85, whereas the importance statements score 4.4 on average. Based on the preliminary results, we observe a discrepancy between how respondents perceive BAIT and what they value. The mean difference between the perception and importance statements appeared significant. We also computed a significance test for the mean difference between both types of statements for each construct. This test showed varying results, by which most differences were significant. Nevertheless, larger sample size is required to induce higher reliability of the conclusions.

The fourth sub-part of the descriptive analysis presents the relationships between the constructs and HMA. Four out of five relationships indicate respondents' perceptions of HMA in BAIT resonate with the underlying theory similarly. As argued in chapter 5, the underlying theory implies that, whenever any of the constructs are being perceived as low, their moral autonomy is being perceived low too, and vice versa. Hence, we recognise a positive relationship between HMA and its subordinate constructs based on most perception figures. This preliminary evidence is not conclusive to confirm or reject the theory. However, the current findings point in the direction of a positive association between the constructs. The positive relationship proves a similarity between social acceptance and ethical acceptability. We argued earlier that when person X has insufficient accessibility to a digital device in the context of decision-making (given that all other conditions are met), the person is considered not autonomous. Now, when the same person does have sufficient access to a digital device, the supposition implies the person is morally autonomous (ceteris paribus). The positive relationship between HMA and the sub-conditions substantiates this hypothesis. This confirmatory (yet insignificant) evidence is promising and should encourage other researchers to elaborate on this concept.

Overall, we may conclude that (1) respondents score high on the constructs of BAIT (on average), (2) their perceptions of HMA in BAIT is relatively positive and (3) there is a slight discrepancy between the perceptions of BAIT and the importance of constructs. BAIT does respect the conditions of HMA according to this sample of end-users. However, due to high data uncertainty, we cannot extract conclusive evidence from the data. There appears a substantial dispersion of data points, which disables further generalisation or conclusions. Therefore, we argue to achieve bigger sample sizes to reach higher validity of data.

Limitations and implications

Sample information showed a vast imbalance in the number of respondents between organisations. Presumably, mainly the perceptions of UMCG surgeons are reflected in the results. Moreover, the computation of the mean difference entails peculiar limitations too. The disclaimer within these results is that the perception contains four statements for each construct, and importance contains one statement. The statistical comparison is not entirely balanced. Additionally, the low number of respondents made it complicated to interpret the results, as it entailed high uncertainty. We could mostly find positive relationships in the relationship plots, but they primarily appeared with relatively high dispersion. Hence, we could not provide conclusive evidence. Although we aimed to identify exciting patterns and relations in the data, more respondents are required to support the arguments.

7.1.5 Improving BAIT

The fifth part of this study entails the guidelines to improve BAIT in terms of HMA. Based on the descriptive analysis and an extensive literature, we formulated recommendations to improve the technology. Moreover, we demonstrate how the survey results can be used to identify points of improvement. This subsection answers the following sub-question:

Research question 5

5. How can DSSs like BAIT be designed so that they better respect human moral autonomy?

We propose recommendations that directly fit the hypothetical constructs of the HMA survey. This proposal is excluding the overarching HMA construct. By doing so, we demonstrate how BAIT can be improved intrinsically. Therefore, we selected scientific articles that showed conceptual overlap with each of the constructs. We did this for two reasons. First, we aim to improve the technology in terms of HMA. As we proposed the conditions of HMA earlier, we only consider those improvements that may enhance HMA. Second, we want to improve those features of the system that we can measure with the HMA survey. Accordingly, the guidelines contain two or three recommendations for each construct. We provided statistical substantiation by using the mean difference between the perception and importance statements. We did this to examine the difference between what respondents perceive and how they value each statement. If the mean difference is statistically significant, it implies the null hypothesis can be rejected (i.e. the difference is not caused by coincidence). Altogether, this methodology enabled us to formulate concise recommendations to improve BAIT in terms of HMA.

To enhance the accessibility of the system, we recommend designing mental models and make sure end-users can make one themselves. The second and third recommendations are related to the first recommendation. The latter two emphasise the provision of complete and sound explanations on the system components. The full set of recommendations are formulated to meet the definition of accessibility by Van den Hoven (Van den Hoven, 1998): "understanding the system as a matter of fact". Hence, this can be done by interactively co-operating with the end-user during the presentation of the model. Additionally, all of the explanations of the model need to be complete and sound. An iterative method supports improving the explanations by measuring the completeness and soundness of each of the concepts.

The improvement of BAIT on tractability can mainly be done by providing human-like reasoning, present a step-by-step computation of the model and solely use familiar terms and justifications to the end-user. Tractability is defined by Van den Hoven (Van den Hoven, 1998) as "the ability to keep track of the system". The recommendations mainly entail a step-wise approach in which the end-users can track the origination of the advice provided by the model. This improvement may encourage putting BAIT into the frame of the usual human-thinking process. Consequently, tractability does include the technical features and the way it is presented to the user. Except for the conceptual solution we propose here, we have not found practical ways to integrate this within BAIT.

Getting high-fidelity human expertise consists of three recommendations: let BAIT perform pre-selection and use thresholds, avoid higher levels of automation and ensure sufficient time for end-user. Using BAIT for a pre-selection is a first step towards the ability to get human advice. After the pre-selection, the end-user would still be able to get advice from peers. The usage of thresholds ensures sufficient time for the end-user by distinguishing cases by complexity. Accordingly, the end-user can comprehend what issues to focus on in the decision-making. Avoidance of higher levels of automation is a topic we touched upon before in the first sub-question. Based on the level of automation scale (see fig. 2.7), we advise not to exceed the level of automation to preserve BAIT to stay a partially autonomous agent. Since higher levels of automation imply fewer tasks are allocated to the human agents, it becomes even harder to get advice from peers during the decision-making. The last recommendation is coherent with the previous one: ensure sufficient time for the end-user. By ensuring adequate time to the end-user, it is possible to get advice from peers. Consequently, we provide these recommendations to provide the ability to the human agent to leave the artificial epistemic niche sufficiently.

To identify and empower human expertise, we recommend considering differences in the expert level and ensuring the end-users independence. Moreover, we recommend identifying different types of knowledge and present this clearly to the end-users. The first recommendation distinguishes between levels of expertise to ensure end-users do not rely exceedingly on BAIT. We assume that decision-makers with higher levels of expertise are more able to evaluate the technology than juniors. More reliance on the technology would then be less harmful to the moral autonomy

of the user. Nevertheless, this contrasts with what philosophers would argue, as they stand for unconditional respect for HMA despite expertise level. The purpose of BAIT is not to replace human intelligence and tasks completely but to support the decision-maker. We, therefore, propose as a second recommendation to distinguish different types of knowledge and acknowledge this. Providing conceptual knowledge could hypothetically increase the understanding of the human agent and hence improve its HMA.

To improve critical questioning and evaluation of the end-user, we recommend informing the end-users on the different types of decision-making and use multiple validation methods to enhance involvement. By informing end-users on the different kinds of decision-making, we aim to induce awareness for the hybrid setting in which humans make decisions with the support of algorithms. They do not replace one another but may instead reinforce if applied correctly. Lastly, we recommend developing prototypes per group of end-users and collectively validate the model in several steps.

Limitations and implications

The last part of this study entails the guidelines for BAIT to improve HMA. We decided to use five theoretical constructs of the framework to formulate the points of improvement. The research is done by searching for best practices of technologies within the literature. We aimed to find potential issues of improvement for BAIT to enhance the HMA of end-users increasingly. As this is a formal proposal for the technology, it has not been validated yet. The recommendations are on a conceptual level and are therefore not tested on their applicability. If further development of BAIT is devoted to these guidelines, it will not directly constitute a maximum degree of HMA. It rather implies a conscious process towards a human-centred approach.

7.2 Reflection on Ethical acceptability and Social acceptance

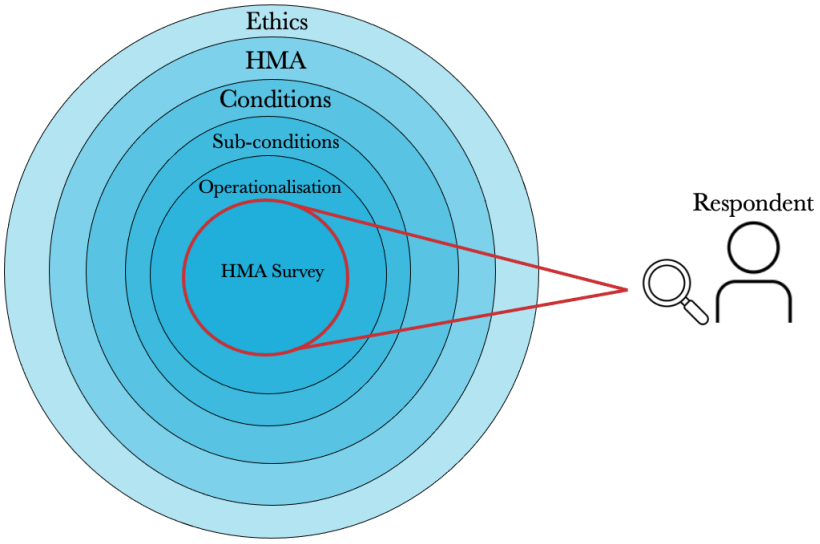
We repeatedly mentioned the concepts of Ethical Acceptability (EA) and Social acceptance (SA) within this thesis. The former entails an ethical reflection based on moral theories, whereas the latter evaluates the degree of acceptance among the concerned group of people. The results of this study involve implications for both concepts, which we reflect upon in this section.

This study used an empirical ethics approach: we used normative concepts with a social approach to study HMA of BAIT. There are numerous ways to conduct an empirical ethics research. As fig. 7.3 shows, we studied merely HMA as sub-part of ethics, in which we extracted conditions and sub-conditions, operationalised them and eventually formulated them in the HMA Survey. This description suggests we started from fundamental ethical theories. On that account, respondents are presented a small portion of statements which we hypothesised to be important. Conversely, we could have decided to first conduct interviews and extract hypothesised (sub-)conditions from the input of potential users. By doing so, we could relate the results to the literature. Nevertheless, we decided to start with the ethical literature, as philosophers have thought thoroughly on the subject of moral autonomy. This enabled us to channel the concept of moral autonomy assuredly into a measurement instrument. Hence, we argue our approach is particularly suitable, as the hypothesised constructs are a compromise of the ethical literature.

The chosen research method ensured a well-grounded theoretical foundation, based on which we constructed the survey. However, this method excludes the input of the respondents in the initial phase. The respondents were only able to express their perceptions of the predetermined constructs. An alternative research method could possibly lead to significantly different constructs. However, the alternative would likely show less overlap with the literature. Moreover, it could lead implicitly to scientific flaws (e.g. incorrect assumption of relations between constructs, insufficient theoretical background, etc.). Both our method and other alternatives have their pros and cons. However, we consider a reliable theoretical foundation to be the most important aspect. Hence, we arguably decided to follow this research method.

The concepts of EA and SA appear mutually throughout this report. The theoretical framework is primarily designated to EA. It clarifies the acceptability of HMA in computerised environments

Figure 7.3: Reflection of empirical ethics approach



and prescribes the determinant conditions of morally autonomous humans. The HMA survey contains both EA and SA aspects. The statements in the initial survey are primarily based on the hypothesised constructs from the HMA framework. However, the survey is validated based on the input of respondents. Although we assumed the statements would converge into the hypothesised constructs, the factor analysis proved the opposite. Hence, the final selection of components increasingly reflects the perception of the respondents. Chapter 6 also contains both concepts of SA and EA. The guidelines are formulated based on a literature review and the descriptive analysis of the survey. To conclude, EA and SA are distinctive concepts, but mutually appear in this thesis.

Another noteworthy point of interest is the contribution of empiricism to a philosophical study. We noticed the empirical part enabled us to translate normative concepts into something we aim to describe pragmatically. However, we emphasise the importance of taking into account the ability of respondents to answer moral statements. The respondents should know about the concerned technology, realise the hypothetical cases in which it is used and understand the implicit normative concepts. Those prerequisites contribute to the validity of the measuring instrument.

The empirical results do not imply that technology should adhere to the general notion of the audience. In contrast, we aim to induce a broader debate on the ethical issues of technological developments. Often, ethical debates are limited to the arena of philosophers. We do not argue that normative affairs should be completely excluded but used to bridge the gap between ethical and empirical realities. By including both distinctive concepts, we are able to find the most crucial points of improvement of the technology.

Even so, a valid and reliable measurement instrument does not suggest to entirely replace moral theories. It is vital to use both empirical methods and ethical theories to enhance both concepts interchangeably. Surveys measure perceptions, and perceptions are fluid in terms of technology, time and context. The constant evaluation of normative ideas is crucial to create suitable measurement instruments.

To conclude, survey studies on moral issues are applicable but can be conducted in versatile ways. Each approach possesses its implications and limitations and should be considered accordingly.

7.3 Scientific contribution

Literature shows that there is a pressing need for further research on perceptions of automated decision-making tools that operate from a hybrid approach (i.e. collaboration between human agent and artificial agent) (Araujo et al., 2020). Moreover, perceptions of heterogeneous groups differ vastly among society, as it is subject to many characteristic features (Lee & Baykal, 2017). However, most studies measure generic perceptions of algorithms on many ethical values (Helberger et al., 2020; Smith, 2018; Dietvorst et al., 2015; Lee & Baykal, 2017; Shin, 2020). Within this study, we solely focused on one particular technology, taking into account one ethical value and

used the expertise of a selective group of decision-makers. Before conducting the empirical study, we thoroughly studied human-computer interaction between BAIT and potential end-users. We found the possible allocation of tasks to enhance moral responsibility among end-users. After we defined the theoretical constructs within the framework (chapter 3), we translated those into measurable variables in a survey study. The forthcoming results on perceptions and the guidelines can be used to study the HMA of BAIT more thoroughly, as explained in section 8.2.2.

Our main scientific contribution is the translation of theoretical (normative) constructs into a survey study, by which we were able to measure the perceptions of potential end-users. By doing so, we compared the perceptions of respondents with the underlying theoretical constructs. We found out that they mostly perceive the theoretical constructs of BAIT similarly to the underlying meaning. Although we did not use a validated and equipped framework to measure the perceptions, we were able to measure respondents' perceptions adequately. By integrating the literature within an empirical study, this research uniquely positions itself in the present-day literature.

7.4 Practical contribution

This research taps into normative and descriptive perspectives to offer Councyl actionable advice to (1) effectively communicate to their clients and (2) improve the technology. A first exploratory study of HMA and the perceptions of HMA provides many insightful points to improve the technology. We provided information on how BAIT can practically be used within organizations, what developers must beware of and how they can respond to a pressing demand for ethical algorithms. Furthermore, we revealed the perceptions of experts on the technology. Accurate solutions can be drawn from this analysis. The guidelines in the latter part of this study are the practical recommendations that can be used to improve the technology towards a human-in-command approach.

An essential practical contribution of this research is the designed survey, which we directly translated from the theoretical framework. The survey is helpful to measure the perceptions of the included constructs and understand the differences over time. The survey is explicitly made for BAIT. However, it can be used to evaluate other technologies too. Researchers may elaborate on the current survey to assess other technologies too.

The guidelines within this research have been formulated based on the studied constructs. The explicit practical recommendations we propose are valuable points to improve the technology in terms of HMA. The recommendations align with ethical theories on human-computer interaction and cater to the international AI ethics guidelines. The recommendations are, by definition, the most practical points we propose to improve BAIT.

Chapter 8

Conclusion

The main goal of this research is to understand the concept of HMA in the context of BAIT, develop a theoretical framework and set up an HMA survey. This chapter presents the main conclusions of this research in section 8.1, the recommendations for the problem owner and scientific field in section 8.2 and a brief reflection on the link with the Engineering and Policy Analysis study program in section 8.3.

8.1 Main Conclusion of Research

Within this research, we aimed to formulate an answer to the following main research question:

Main research question

How can we define, operationalise and measure the (perceived) human moral autonomy of decision support systems like behavioural artificial intelligence technology?

The answer to the main research question is multi-faceted, as it contains three verbs to conceptualise HMA:

1. In what ways does BAIT support human decision-making in organisations, and how does that affect task allocations between the human and the DSS?
2. What is a philosophical definition of human moral autonomy in the context of a DSS like BAIT?
3. How can we measure the degree to which DSSs like BAIT respect human moral autonomy using a survey?
4. To what extent does BAIT respect the conditions for HMA according to end-users?
   (a) What importance do end-users ascribe to the philosophical conditions for HMA?
   (b) Does BAIT respect the philosophical conditions for HMA in the perception of end-users?
5. How can DSSs like BAIT be designed so that they better respect human moral autonomy?

The first sub-question describes the supportive role of the technology and the tasks it can execute. BAIT codifies expert knowledge utilising Discrete Choice Modelling (DCM) to present advice to decision-makers. This advice, provided by the system, is solely for supportive purposes to improve the decision-making. Depending on the automation level, BAIT is at least able to suggest one choice alternative. More advanced automation levels imply the system independently executes tasks, in which the human merely evaluates the decision. This description is what we refer to as 'partial agent autonomy'. Thus, the task allocation changes with increasing automation levels. The human decision-maker increasingly has the task of evaluating the decisions rather than making decisions by themselves with rising automation levels.

The second sub-question philosophically defines human moral autonomy (HMA) in the context of BAIT. This research characterises moral autonomy by determining the conditions under

which we consider humans to be morally autonomous. The definition of HMA within this research is: "A person is morally autonomous if and only if he bears the responsibility and authority for moral supervision on the situation and the decision support system". The conditions under which we consider humans to be morally autonomous are: (1) inscrutinizability, (2) pressure condition, (3) error condition and (4) critical questioning and evaluation. Hence, the definition of HMA stands by satisfaction of these conditions. The conditions are divided in sub-conditions and operationalised accordingly. We defined a theoretical framework on moral autonomy in computerised environments to translate the normative constructs into measurable variables. The following six theoretical constructs are more profoundly studied within this research: (1) HMA, (2) accessibility, (3) tractability, (4) getting high-fidelity human expertise, (5) identify & empower human expertise and (6) critical questioning and evaluation. HMA is the overarching construct by which the latter five are subordinate.

The third sub-question aims to set up the HMA survey (i.e. the measurement instrument) by which we are able to estimate the perceived HMA of end-users. The survey reflects all of the constructs as clarified above. The first validation is the pilot survey. The pilot survey entails interviews with respondents from the concerned organisations. We concluded that respondents were capable to answer the statements of our survey. The statistical validation of the output data, utilising factor analysis, resulted in an adjusted combination of components that reflect differing constructs. The adjusted version of the survey contains the following components: (1) Choice reflection, (2) Autonomy, (3) Boundless decision-making, (4) Intelligibility, (5) Explainability (6) Knowledge interchange. The validated version contains primarily statements of the overarching HMA construct and the critical questioning and accessibility sub-conditions. Hence, it contains disproportionally more statements of those constructs than the other sub-conditions. This imbalance implies the validated HMA survey mainly measures those constructs to evaluate the perception of moral autonomy. To conclude, we are able to measure the degree to which BAIT respects HMA using a survey. We proved the feasibility of translating normative concepts into measurable variables. However, further development of the survey is paramount to ensure its validity.

The fourth sub-question analyses the responses of the survey to evaluate the perceptions. The survey contains perception and importance statements. The former reveals respondents' experience regarding the technology, whereas the latter directly measures social acceptance concerning the sub-conditions. This enabled us to study the difference between how people perceive their moral autonomy and the importance they ascribe to each of the predetermined constructs. Respondents scored on average high for both types of statements (3.85 perception and 4.4 importance on average).

We studied the mean differences between both statements for each construct to identify the divergence between both types. Most constructs demonstrated a significant difference. This insight implies respondents relatively scored higher on the importance they ascribe to constructs than their perception within the context of BAIT. However, this does not suggest BAIT fails in terms of the sub-conditions. In contrast, it may support to identify points of improvement.

Moreover, we evaluated the relationships between all sub-conditions and HMA. The relationships predominantly showed a positive relationship. This demonstrates an approximate similarity between the underlying hypotheses of the constructs and the perceptions of respondents. However, the little number of respondents caused a considerable data uncertainty. Hence, we cannot draw conclusive evidence from the results.

The fifth sub-question formulates points of improvement for BAIT in terms of the selected theoretical constructs. The proposal is based on best practices from the literature. Additionally, we substantiated our proposal with the analysis of perceptions from the previous sub-question. The recommendations are specifically formulated to improve BAIT to better respect moral autonomy of users. The recommendations enable to measure the differences in perception over time and enhance end-users' moral autonomy.

The conclusive answer to the main research question is three-fold. Firstly, we were able to define BAIT by comparing its characteristics with conventional systems. After that, we described HMA in the context of BAIT. On that account, we operationalised the (sub-)conditions into the so-called HMA framework. Subsequently, we measured the perceptions of potential end-users with the HMA survey, which directly reflects the sub-conditions. By analysing the data, we were able to illustrate the perceptions quantitatively. We demonstrated the feasibility of translating normative

concepts into an empirical study. Albeit the number of respondents was too low to draw conclusive evidence, we managed to obtain insightful output.

8.2 Recommendations

This section entails two types of recommendations; the first part addresses recommendations for the problem owner, Councyl, and the second part provides recommendations for further scientific research.

8.2.1 Recommendations for Councyl

The first recommendation to Councyl, the problem owner, is to implement the recommendations in the proposed guidelines of this study. The guidelines entail technical improvements to enable HMA and conceptual propositions to put BAIT into a narrative of human-centred technology. This research generally puts forward the importance of how technologies can be built to support humans in decision-making adequately. We, therefore, recommend Councyl, as argued before, not to exceed the automation levels for the sake of efficiency. BAIT is able to compete against many other algorithms (often black-box models) if it stays true to its current features.The transition towards algorithmic decision-making (ADM) demands a procedure where end-users are guided and taught how to interact with such systems. Hence, algorithmic decision-making is more than just mathematical computations. It is an advanced interaction between humans and computers; the latter must always serve the needs and respect the values of the former. By doing so, Councyl may fill in the societal and market gap by developing a technology that accurately puts its focus on the end-user.

Finally, we recommend Councyl improve the technology by constantly evaluating its model with its clients in multiple ways. Firstly, we recommend using the survey we made within this research to measure their users' perceptions continuously. By that means, Councyl will be able to evaluate the perceptions of HMA and subsequently improve their technology. Secondly, we recommend developing and measure user evaluations on certain technical features using contemporary methodologies. By doing so, Councyl can evaluate and improve the technical features of BAIT. More importantly, they would enable themselves to compare BAIT with various other algorithms and concretely point out the differences and similarities with other technologies. Consequently, Councyl is able to distinguish BAIT from other technologies more precisely.

8.2.2 Recommendations for future scientific research

The plurality of definitions of human moral autonomy required studying a multitude of ideas by leading philosophers. We found out that human moral autonomy constitutes a multi-dimensional definition, appearing in many sub-domains of ethics and touching upon many conditions. Especially the conceptual synergy between human moral autonomy and digital systems resulted in numerous characterisations of how human moral autonomy may formalise in the digital environments. For practical reasons, we had to limit ourselves to a limited number of scientific papers. As mentioned earlier, this research concretely approaches from one theoretical approach, which has been quantified through a survey. However, we always aimed to put this in the context of other theoretical perspectives. Future research may be conducted from other theories to understand the manifestation of human moral autonomy from different theoretical perspectives.

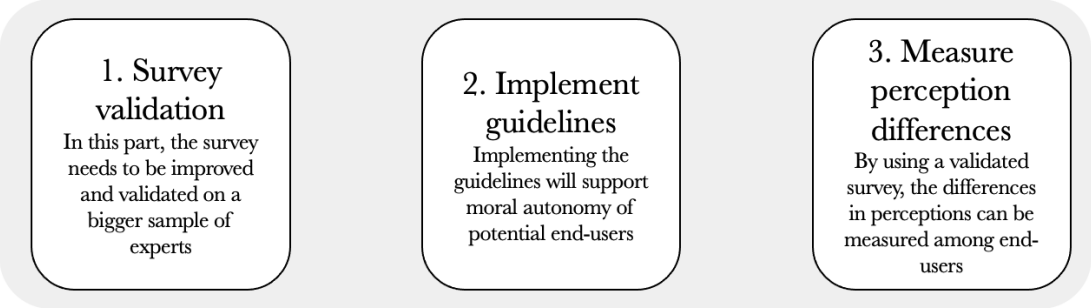


Figure 8.1: Scientific recommendations

There are also recommendations on how one could build further on the premise of this study. Due to time constraints, it was impossible to validate the survey and evaluate the proposed guidelines. Therefore, we formulate three main recommendations to build further on this study, as shown in fig. 8.1. First, we recommend validating the survey for further applications. To realise this, one requires a more considerable sample size to reach statistical significance. Not only does the sample require a higher number of respondents, but it is also paramount to validate it in a later development stage of BAIT. The technology is still in the developing phase and has not been implemented in an organisation yet. The survey can be easier to validate potential respondents when they are more familiar with the technology. Subsequently, we advise implementing the recommendations in the proposed guidelines in chapter 6. This activity leads to a hypothesised increased level of moral autonomy for end-users of BAIT. After validating both the survey and having made progress with implementing the guidelines, one could study the perception differences of end-users on the various theoretical constructs. It is therefore crucial that both recommendations mentioned above have been conducted successfully to continue with this recommendation. This step is essential to measure the differences in perceptions.

8.3 Link with EPA program

The Engineering and Policy Analysis (EPA) programme integrates technology and society with an analytical approach. The interdisciplinary nature of this study provides knowledge and the means to analyse problems, model and simulate dynamic systems and assess solutions. This thesis combines the societal aspect with the technological part by studying the effects of a novel technology on the moral autonomy of potential users. We combined ethical, empirical, and technical perspectives to answer complex problems occurring on societal and company levels. By synthesising both conceptual aspects, we aimed to provide an interdisciplinary solution to the rapid upswing of AI technologies.

References

Abraham, A. (2005, 7). Rule-Based Expert Systems. In *Handbook of measuring system design*. Chichester, UK: John Wiley & Sons, Ltd. Retrieved from http://doi.wiley.com/10.1002/0471497398.mm422 doi: 10.1002/0471497398.mm422

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

AI HLEG. (2019). Ethics guidelines for trustworthy AI. *European Commission*, 1–39.

Alexander, G. L. (2006). Issues of trust and ethics in computerized clinical decision support systems. *Nursing Administration Quarterly*, *30*(1), 21–29. doi: 10.1097/00006216-200601000-00005

Anderson, L. W., Bloom, B. S., et al. (2001). *A taxonomy for learning, teaching, and assessing: A revision of bloom's taxonomy of educational objectives*. Longman,.

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI and Society*, *35*(3), 611–623. Retrieved from https://doi.org/10.1007/s00146-019-00931-w doi: 10.1007/s00146-019-00931-w

Arnott, D., & Pervan, G. (2005). A critical analysis of decision support systems research. *Journal of Information Technology*, *20*(2), 67–87. doi: 10.1057/palgrave.jit.2000035

Aronson, J. E., Liang, T.-P., & MacCarthy, R. V. (2005). *Decision support systems and intelligent systems* (Vol. 4). Pearson Prentice-Hall Upper Saddle River, NJ, USA:.

Arrow, K. J. (1974). *The limits of organization*. WW Norton & Company.

Baalen, S., Boon, M., & Verhoef, P. (2021, 2). From clinical decision support to clinical reasoning support systems. *Journal of Evaluation in Clinical Practice*, jep.13541. Retrieved from https://onlinelibrary.wiley.com/doi/10.1111/jep.13541 doi: 10.1111/jep.13541

Bader, V., & Kaiser, S. (2019). Algorithmic decision-making? the user interface and its role for human involvement in decisions supported by artificial intelligence. *Organization*, *26*(5), 655–672.

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach* (Vol. 904). John Wiley & Sons.

Ben-Akiva, M., & Lerman, S. R. (2018). *Discrete choice analysis: theory and application to travel demand*. Transportation Studies.

Ben-Akiva, M. E., McFadden, D., Train, K., et al. (2019). *Foundations of stated preference elicitation: Consumer behavior and choice-based conjoint analysis*. Now.

Berner, E. S., & La Lande, T. J. (2007). Overview of clinical decision support systems. In *Clinical decision support systems* (pp. 3–22). Springer.

Bloom, B. S., et al. (1956). Taxonomy of educational objectives. vol. 1: Cognitive domain. *New York: McKay*, *20*, 24.

Borenstein, D. (1998). Towards a practical method to validate decision support systems. *Decision Support Systems*, *23*(3), 227–239.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 1165–1188.

Chiodo, S. (2021). Human autonomy, technological automation (and reverse). *AI & SOCIETY*, *1*, 3. Retrieved from https://doi.org/10.1007/s00146-021-01149-5 doi: 10.1007/s00146-021-01149-5

Chorus, C. G. (2015, 9). Models of moral decision making: Literature review and research agenda for discrete choice analysis. *Journal of Choice Modelling*, *16*, 69–85. doi: 10.1016/j.jocm.2015.08.001

Chung, K., Boutaba, R., & Hariri, S. (2014). Recent trends in digital convergence information system. *Wireless Personal Communications*, *79*(4), 2409–2413.

Creswell, J. W. (1999). Mixed-method research: Introduction and application. In *Handbook of educational policy* (pp. 455–472). Elsevier.

Cummings, M. (2006). Automation and Accountability in Decision Support System Interface Design M.L. Cummings 1 Massachusetts Institute of Technology. *The Journal of Technology Studies*, *32*(1), 23–31.

Dattachaudhuri, A., Biswas, S., Sarkar, S., & Boruah, A. N. (2020). Transparent decision support system for credit risk evaluation: An automated credit approval system. In *2020 ieee-hydcon* (pp. 1–5).

Dearden, R. F. (1972). Autonomy and education. *Education and the development of reason*, *58*, 75.

Deloitte. (2021). https://www2.deloitte.com/nl/nl.html. (Accessed: 2021-04-30)

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. Retrieved from http://doi.apa.org/getdoi.cfm?doi=10.1037/xge0000033 doi: 10.1037/xge0000033

Downie, R. S., & Telfer, E. (1971). Autonomy. *Philosophy*, *46*(178), 293–301.

Dworkin, G. (1981). Moral Autonomy. *Common-Sense Morality and Consequentialism*, 23–34. doi: 10.4324/9781003049265-3

Dworkin, G. (2015, 1). The nature of autonomy†. *Nordic Journal of Studies in Educational Policy*, *2015*(2). Retrieved from https://www.tandfonline.com/action/journalInformation?journalCode=znst20 doi: 10.3402/nstep.v1.28479

Edwards, W. (1954). The theory of decision making. *Psychological bulletin*, *51*(4), 380.

Fahlman, S. E. (2011). Using scone's multiple-context mechanism to emulate human-like reasoning. In *2011 aaai fall symposium series*.

Feinberg, J. (1982). Autonomy, sovereignty, and privacy: Moral ideals in the constitution. *Notre Dame L. Rev.*, *58*, 445.

Floridi, L. (2019, 6). *Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical* (Vol. 32) (No. 2). Springer Netherlands. Retrieved from https://doi.org/10.1007/s13347-019-00354-x doi: 10.1007/s13347-019-00354-x

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*(1), 1–15. doi: 10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... Vayena, E. (2018, 12). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. Retrieved from https://doi.org/10.1007/s11023-018-9482-5 doi: 10.1007/s11023-018-9482-5

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. In *Minds and machines* (pp. 349–379). doi: 10.4324/9781003074991-30

Frankena, W. K. (1939). The naturalistic fallacy. *Mind*, *48*(192), 464–477.

Fricker, R. D., Kulzy, W. W., & Appleget, J. A. (2012). From Data to Information: Using Factor Analysis with Survey Data. *Phalanx*, *45*(4), 30–34.

Gass, S. I. (1983). Decision-aiding models: validation, assessment, and related issues for policy analysis. *Operations Research*, *31*(4), 603–631.

Goodman, B., & Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, *38*(3), 50–57.

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, *23*(2), 101–124. Retrieved from https://www.tandfonline.com/action/journalInformation?journalCode=hpli20 doi: 10.1080/1047840X.2012.651387

Grimm, P. (2010). Pretesting a Questionnaire. *Wiley International Encyclopedia of Marketing*, 2010. doi: 10.1002/9781444316568.wiem02051

Guerlain, S., Brown, D. E., & Mastrangelo, C. (2000). Intelligent decision support systems. In *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics.'cybernetics evolving to systems, humans, organizations, and their complex interactions'(cat. no. 0* (Vol. 3, pp. 1934–1938).

Gunning, D. (2016). Explainable artificial intelligence (xai) darpa-baa-16-53. *Defense Advanced Research Projects Agency*.

Hagras, H. (2018, 9). Toward Human-Understandable, Explainable AI. *Computer*, *51*(9), 28–36. doi: 10.1109/MC.2018.3620965

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., ... others (2020). Transparency and reproducibility in artificial intelligence. *Nature*, *586*(7829), E14–E16.

Harman, H. H. (1976). *Modern factor analysis*. University of Chicago press.

Helberger, N., Araujo, T., & de Vreese, C. H. (2020, 11). Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law and Secu-*

rity Review, *39*, 105456. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0267364920300613 doi: 10.1016/j.clsr.2020.105456

Hensher, D. A., & Assoiate, M. B. (1993). *Using Stated Response Choice Data to Enrich Revealed Preference Discrete Choice Models* (Vol. 4; Tech. Rep.).

Hensher, D. A., Greene, W. H., & Chorus, C. G. (2013). Random regret minimization or random utility maximization: an exploratory analysis in the context of automobile fuel choice. *Journal of Advanced Transportation*, *47*(7), 667–678.

Hensher, D. A., Greene, W. H., & Ho, C. Q. (2016). Random regret minimization and random utility maximization in the presence of preference heterogeneity: an empirical contrast. *Journal of Transportation Engineering*, *142*(4), 04016009.

Hensher, D. A., Rose, J. M., & Collins, A. T. (2011). Identifying commuter preferences for existing modes and a proposed metro in sydney, australia with special reference to crowding. *Public Transport*, *3*(2), 109–147.

Hughes, K. K., & Young, W. B. (1990). The relationship between task complexity and decision-making consistency. *Research in nursing & health*, *13*(3), 189–197.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kant, I. (1785). Groundwork of the metaphysics of morals.(1785).

Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015, 3). Principles of Explanatory Debugging to personalize interactive machine learning. In *International conference on intelligent user interfaces, proceedings iui* (Vol. 2015-January, pp. 126–137). Association for Computing Machinery. Retrieved from http://dx.doi.org/10.1145/2678025.2701399 doi: 10.1145/2678025.2701399

Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012). Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1–10).

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 ieee symposium on visual languages and human centric computing* (pp. 3–10).

Lee, M. K., & Baykal, S. (2017, 02). Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. , 1035-1048. doi: 10.1145/2998181.2998230

Lim, B. Y., & Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on ubiquitous computing* (pp. 195–204).

Longhurst, R. (2003). Semi-structured interviews and focus groups. *Key methods in geography*, *3*(2), 143–156.

Lucas, J. R. (1966). *The principles of politics*. Clarendon Press Oxford.

Marakas, G. (1999). *Decision support systems in 21st century–us edition.* Upper Saddle River, London: Prentice Hall.

McFadden, D. (1986). The choice theory approach to market research. *Marketing science*, *5*(4), 275–297.

McFadden, D. (2001). Economic choices. *American economic review*, *91*(3), 351–378.

McFadden, D., et al. (1973). Conditional logit analysis of qualitative choice behavior.

Merino-Castello, A. (2003). Eliciting consumers preferences using stated preference discrete choice models: contingent ranking versus choice experiment. *UPF economics and business working paper*(705).

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016, 12). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. Retrieved from http://journals.sagepub.com/doi/10.1177/2053951716679679 doi: 10.1177/2053951716679679

Mökander, J., & Floridi, L. (2021, 2). *Ethics-Based Auditing to Develop Trustworthy AI.* Springer Science and Business Media B.V. Retrieved from https://doi.org/10.1007/s11023-021-09557-8 doi: 10.1007/s11023-021-09557-8

Molin, E. J., & Timmermans, H. J. (2010). Context dependent stated choice experiments: The case of train egress mode choice. *Journal of Choice Modelling*, *3*(3), 39–56.

Montani, S., & Striani, M. (2019, 8). *Artificial Intelligence in Clinical Decision Support: a Focused Literature Survey* (Vol. 28) (No. 1). NLM (Medline). Retrieved from /pmc/articles/PMC6697510//pmc/articles/PMC6697510/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6697510/ doi: 10.1055/s-0039-1677911

Murphy, J. J., Allen, P. G., Stevens, T. H., & Weatherhead, D. (2005). A meta-analysis of

hypothetical bias in stated preference valuation. *Environmental and Resource Economics*, *30*(3), 313–325.

Musschenga, A. W., & Musschenga, A. W. (2005). Empirical Ethics, Context-Sensitivity, and Contextualism. *Journal of Medicine and Philosophy*, *30*(05), 1–27. Retrieved from https://academic.oup.com/jmp/article/30/5/467/922582 doi: 10.1080/03605310500253030

Nelson Ford, F. (1985, 1). Decision support systems and expert systems: A comparison. *Information and Management*, *8*(1), 21–26. doi: 10.1016/0378-7206(85)90066-7

Nissenbaum, H. (1996). Accountability in a computerized society. *Science and Engineering Ethics*, *2*(1), 25–42. Retrieved from https://link.springer.com/article/10.1007/BF02639315 doi: 10.1007/BF02639315

Nolan, J. R. (1998). An intelligent system for case review and risk assessment in social services. *AI Magazine*, *19*(1), 39–39.

Norman, D. A. (1983). Some observations on mental models. *Mental models*, *7*(112), 7–14.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., ... Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3), 1–14. doi: 10.1002/widm.1356

O, T. A., McNeese, N. J., University, C., Carolina, S., & Barron, A. (2020). *Human-Autonomy Teaming: A Review and Analysis of the Empirical Literature* (Vol. 00; Tech. Rep. No. 01).

O'Leary, T. J., Goul, M., Moffitt, K. E., & Radwan, A. E. (1990). Validating expert systems. *IEEE Computer Architecture Letters*, *5*(03), 51–58.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2020). Human–autonomy teaming: A review and analysis of the empirical literature. *Human Factors*, 0018720820960865.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, *30*(3), 286–297. doi: 10.1109/3468.844354

Payne, J. W., Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge university press.

Pearl, J., et al. (2009). Causal inference in statistics: An overview. *Statistics surveys*, *3*, 96–146.

Pérez-Vicente, S., & Ruiz, M. E. (2009). Descriptive statistics. *Allergologia et immunopathologia*, *37*(6), 314–320.

Prentzas, J., & Hatzilygeroudis, I. (2007). Categorizing approaches combining rule-based and case-based reasoning. *Expert Systems*, *24*(2), 97–122.

Price, M., Walker, S., & Wiley, W. (2018). The machine beneath: Implications of artificial intelligence in strategic decision making. *Prism*, *7*(4), 92–105.

Rawls, J. (1971). *A theory of justice*. Harvard university press.

Richards, J. C., & Schmidt, R. W. (2002). *Longman dictionary of language teaching and applied linguistics*. Routledge.

Ritchie, J., & Lewis, J. (2003). *QUALITATIVE RESEARCH PRACTICE A Guide for Social Science Students and Researchers Edited by* (Tech. Rep.).

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of educational psychology*, *91*(1), 175.

Rogers, Y., Sharp, H., & Preece, J. (2011). *Interaction design: beyond human-computer interaction*. John Wiley & Sons.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. Retrieved from http://dx.doi.org/10.1038/s42256-019-0048-x doi: 10.1038/s42256-019-0048-x

Salisbury, M. (2019). When computers advise us: How to represent the types of knowledge users seek for expert advice. *Computer*, *52*(9), 44–51.

Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica*, *15*(60), 243–253.

Santoni de Sio, F., & van den Hoven, J. (2018, 2). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI*, *5*(FEB), 28. Retrieved from http://journal.frontiersin.org/article/10.3389/frobt.2018.00015/full doi: 10.3389/frobt.2018.00015

Sarter, N., & Woods, D. D. (1994). Decomposing automation: Autonomy, authority, observability and perceived animacy. In *First automation technology and human performance conference* (pp. 22–26).

Sarter, N. B., & Schroeder, B. (2001, 9). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, *43*(4), 573–583.

Retrieved from https://journals.sagepub.com/doi/abs/10.1518/001872001775870403 doi: 10.1518/001872001775870403
Scanlon, T. (1972). *A Theory of Freedom of Expression* (Vol. 1; Tech. Rep. No. 2).
Sen, A., & Biswas, G. (1985, 9). Decision support systems: An expert systems approach. *Decision Support Systems*, *1*(3), 197–204. doi: 10.1016/0167-9236(85)90239-8
Shin, D. (2020, 10). User Perceptions of Algorithmic Decisions in the Personalized AI System:Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541–565. Retrieved from https://www.tandfonline.com/doi/full/10.1080/08838151.2020.1843357 doi: 10.1080/08838151.2020.1843357
Shollo, A., & Kautz, K. (2010). Towards an understanding of business intelligence.
Shortliffe, E. H., & Sepúlveda, M. J. (2018). Clinical decision support in the era of artificial intelligence. *Jama*, *320*(21), 2199–2200.
Sifringer, B., Lurkin, V., & Alahi, A. (2020). Enhancing discrete choice models with representation learning. *Transportation Research Part B: Methodological*, *140*, 236–261. Retrieved from https://doi.org/10.1016/j.trb.2020.08.006 doi: 10.1016/j.trb.2020.08.006
Smith, A. (2018). *Public Attitudes Toward Computer Algorithms — Pew Research Center.* Retrieved from https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/
Spiegelhalter, D. J., & Knill-Jones, R. P. (1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology. *Journal of the Royal Statistical Society: Series A (General)*, *147*(1), 35–58.
Sprague, R. H. (1980). A framework for the development of decision support systems. *MIS Quarterly: Management Information Systems*, *4*(4), 1–26. doi: 10.2307/248957
Sternberg, R. J. (2012, 12). A Model for Ethical Reasoning. *Review of General Psychology*, *16*(4), 319–326. Retrieved from http://journals.sagepub.com/doi/10.1037/a0027854 doi: 10.1037/a0027854
Swait, J., & Ben-Akiva, M. (1987). Incorporating random constraints in discrete models of choice set generation. *Transportation Research Part B: Methodological*, *21*(2), 91–102.
Taebi, B. (2017). Bridging the Gap between Social Acceptance and Ethical Acceptability. *Risk Analysis*, *37*(10), 1817–1827. doi: 10.1111/risa.12734
Tan, T. Z., Ng, G. S., & Quek, C. (2008). Improving tractability of clinical decision support system. In *2008 ieee international joint conference on neural networks (ieee world congress on computational intelligence)* (pp. 1997–2002).
ten Broeke, A. (2020). A new approach to artificial intelligence for decision support. (September), 1–7.
ten Broeke, A., Hulscher, J., Heyning, N., Kooi, E., & Chorus, C. (2021). BAIT: A New Medical Decision Support Technology Based on Discrete Choice Theory. *Medical Decision Making*, 1–6. doi: 10.1177/0272989X211001320
Tjoa, E., & Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, *14*(8), 1–21. doi: 10.1109/tnnls.2020.3027314
Trost, J. E. (1986). Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies. *Qualitative sociology*, *9*(1), 54–57.
Turban, E., & Watkins, P. R. (1986). Integrating expert systems and decision support systems. *MIS Quarterly: Management Information Systems*, *10*(2), 121–136. doi: 10.2307/249031
UCLA. (2021). *Principal Components (PCA) and Exploratory Factor Analysis (EFA) with SPSS.* Retrieved from https://stats.idre.ucla.edu/spss/seminars/efa-spss/#s1
OLVG. (2021). https://www.olvg.nl/. (Accessed: 2021-04-30)
UMCG. (2021). https://www.umcg.nl. (Accessed: 2021-04-30)
Uricchio, V. F., Giordano, R., & Lopez, N. (2004). A fuzzy knowledge-based decision support system for groundwater pollution risk evaluation. *Journal of environmental management*, *73*(3), 189–197.
Van den Hoven, J. (1998). Moral Responsibility, Public Office and Information Technology. *Public administration in an information age : a handbook*, 579.
van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404. Retrieved from https://doi.org/10.1016/j.artint.2020.103404 doi: 10.1016/j.artint.2020.103404
Waa, J. V. D., & Diggelen, J. V. (2020). *Allocation of Moral Decision-Making in Human-Agent*

Teams : (Vol. 2). Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-030-49183-3_16 doi: 10.1007/978-3-030-49183-3

Walker, J. L., Wang, Y., Thorhauge, M., & Ben-Akiva, M. (2015). D-efficient or deficient. In *A robustness analysis of stated choice experimental designs, presented at the 94th annual meeting of the transportation research board, washington, dc.*

Walliser, J. C., de Visser, E. J., Wiese, E., & Shaw, T. H. (2019). Team Structure and Team Building Improve Human–Machine Teaming With Autonomous Agents. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 258–278. doi: 10.1177/1555343419867563

Weiss, S. M., & Kulikowski, C. A. (1991). *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems.* Morgan Kaufmann Publishers Inc.

Wierzynski, C. (2018). The challenges and opportunities of explainable ai. *Intel. com*, *12*.

Wolff, R. P. (1998). *In defense of anarchism.* Univ of California Press.

Woods, D. D. (1996). Decomposing automation: Apparent simplicity, real complexity. *Automation and human performance: Theory and applications*, 3–17.

Wu, H., & Leung, S.-O. (2017). Can likert scales be treated as interval scales?—a simulation study. *Journal of Social Service Research*, *43*(4), 527–532.

Wynne, K. T., & Lyons, J. B. (2018). An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, *19*(3), 353–374. Retrieved from https://doi.org/10.1080/1463922X.2016.1260181 doi: 10.1080/1463922X.2016.1260181

Yong, A. G., Pearce, S., et al. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, *9*(2), 79–94.

Zeleznikow, J., & Nolan, J. R. (2001). Using soft computing to build real world intelligent decision support systems in uncertain domains. *Decision Support Systems*, *31*(2), 263–285.

Zussman, R. (1992). *Intensive care: Medical ethics and the medical profession.* University of Chicago Press.

Appendix A - Pilot survey

The goal of the transcriptions is to perform the pilot survey/pretesting of the questionnaire. Based on the recommendations of Grimm (2010) the questions have been formulated and structured accordingly as shown in fig. 2. The answers will be used as a first validation step for the survey to ensure the quality before deployment. Usually a pilot survey is conducted by supplying the concept questionnaire to the appointed group. For pragmatic reasons we decided to perform an interview in which various questions are asked about the different components and analyse the answers to determine the degree to which respondents are familiar with the terminology.

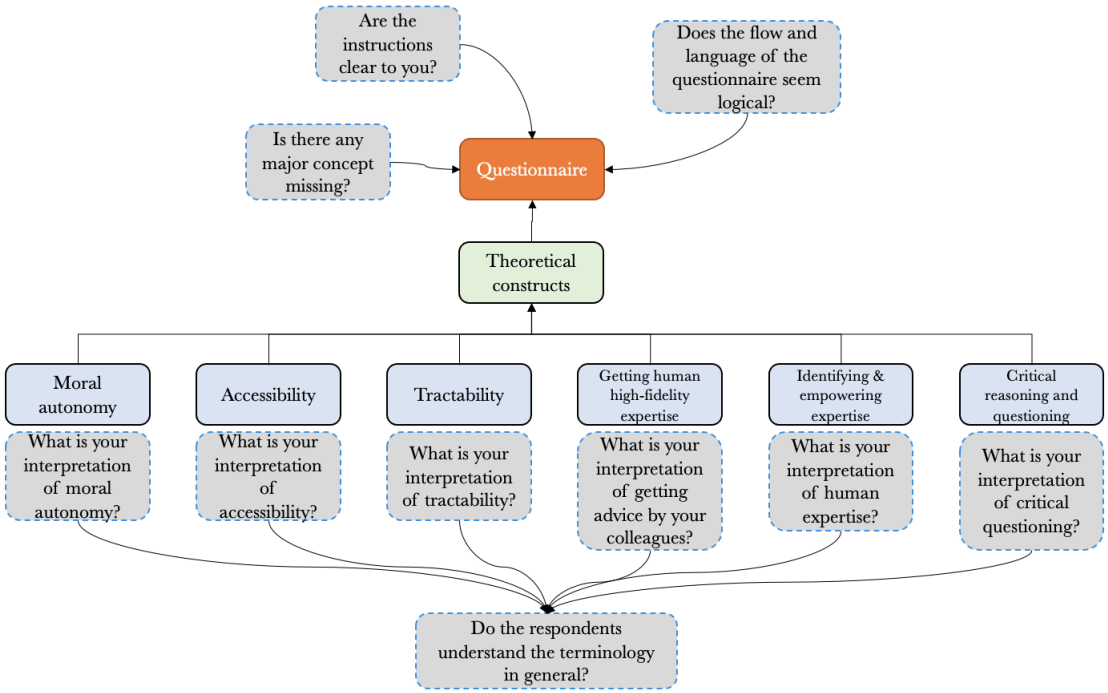


Figure 2: Pilot interview

Deloitte

This interview is held with three key persons who collaborated with Councyl to model BAIT for a specific decision-making. Due to practical reasons, it was not possible to cover all components.

1. **What are your main feelings about the collaboration with Councyl and the results of BAIT?**
   We started with much excitement and found the process genuinely valuable. It mostly provided us insights on how we made decisions in an objective way. The think-process, especially during the criteria inventarisation and choice experiment, obliged us to come up with and judge the importance of criteria. Eventually, it made explicit how we make choices regarding public tenders.

2. **What did the results of the model tell you in general?**
   Except for providing us much insight on how we make our decisions, it mostly did confirm our expectations. In that way, we did not get surprised by the results, which is not necessarily a bad outcome. The awareness it created in our decision-making was already beneficial for our own understanding.

3. **How do you perceive the moral autonomy of potential users within your organisation specifically for this decision-making?**
   For this decision-making, specifically, it is paramount that our colleagues have so-called "industry sensitivity". We believe that this type of sensitivity gets affected when humans rely too much on it. Hence, we believe more in the employment of such models as an actual decision supportive mechanism, rather than for automation purposes. Obviously we argue this solely for this decision-making, as we highly value active thinking of our colleagues. In

contrast we see for example that some of our clients could use a model like this to objectify their decision-making (e.g., when it comes to the relation between the government and citizens).

4. **What hypothetical consequences does the model have for human expertise within your organisation on this decision-making?**
   For less-experienced colleagues of us it would be beneficial to accelerate their learning curve for this decision-making. We believe that it will improve their understanding by proving how certain variables affect the decision. Moreover, it provides insights on the strategy of various colleagues, e.g. someone could be excessively opportunistic in his/her decision to stake on one tender at the expense of another. This is especially interesting to identify different strategies and philosophies among our colleagues.

5. **What are the effects of the type of decision on the way you would use the model?**
   Depending on the impact of the decision, you could use it more or less. Coolblue (company, red.) uses a strategy of abundantly not classifying their clients as fraudulent. They do this deliberately to satisfy their clients, knowing that they possibly miss out on some of the fraudsters. This decision has only consequences for their own business operations and is set up conform their strategy.

6. **Do you feel you can be held responsible when using BAIT more intensively?**
   We believe we do not lose our moral responsibility when we would use BAIT for this decision-making in the context we work. In fact, we think that if we would notice that the model confirms our own predictions on a decision every time, we could even omit the model in such hypothetical case. But, that is only the case when you actively have thought about a particular tender beforehand. If we feel comfortable with the model, we could hypothetically use it for pre-selections without thoroughly reading through each tender. In that sense, we would possibly lose some of our moral responsibility. But, it has low impacts given the stage (pre-selection) and the confidence we would have in the model.

7. **To what extent is it important to independently search for additional information outside the model?**
   For this decision-making we find it important to search for different types of information from various sources. All of our colleagues have their own expertise on various domains and everyone brings his own value to the table in finding the suitable tenders for the correct price. It would not be smart to only rely on a limited number of criteria and leave out all nuances.

OLVG

This interview is held with an intensivist from the intensive care department at AUMC. Councyl and OLVG set up a cooperation to make a choice model for the purpose of admitting covid-19 patients on the IC. Due to practical reasons, it was not possible to cover all components.

1. **What are your main feelings about the collaboration with Councyl and the results of BAIT?**
   We set up the model relatively quickly, but in such pandemic things cannot go fast enough. Especially when the technology is aimed to be used for tackling that exact problem that caused the pandemic. Nevertheless, I see many opportunities in such technologies and was satisfied with the collaborative project. In fact, we operate with more AI-technologies to improve our services to our patients. BAIT is rather uncommon, as most of our innovations are based on big data and machine learning.

2. **What did the results of the model tell you in general?**
   The predictive values were relatively high, explaining 80% of the behaviour of the group of intensivists. We never made the criteria explicit before, so I did not know what to expect in the first place. To me personally, it proved that vaguely defined criteria (beforehand) could practically be translated to concrete variables showing the desired results. This was an eye-opener to me. Moreover, it proved that with an approximate number of 25 choice sets it was possible to obtain sufficient information to build such successful model.

3. **How do you perceive the moral autonomy of potential users within your organisation specifically for this decision-making?**
   For our decision-making it would be used solely for decision supportive purposes. Hence we

would not perceive it as the leading entity decision-maker within this department in the near future. The model only indicates how my colleagues would judge the case given the input. Moreover, it does not include everything (i.e. randomness) and hence one should never rely completely on such models. Within our organisation we are morally always responsible for the decisions we make, such decision support systems are only of supportive functions (as the name already says). It serves as another entity within our team, but we would not give it the same rights and obligations as any member of this organisation.

4. **What hypothetical consequences does the model have for human expertise within your organisation on this decision-making?**
   The assessments we conduct for this decision-making shows an overlap with the model. Therefore, we only use the model for introspective purposes. We as intensivists are trained to identify patterns by ourselves quickly. Within this context the model serves only a supportive function. This function will maintain if you ask me, but I am looking forward to what the future will bring us, also for other domains. The self-learning aspect of such models are extremely important, as rapid developments in healthcare change our policies too. New treatments could, for example, lead to different trade-offs and decisions (e.g. new lung cancer treatments could determine whether or not to admit patient X to the IC). The model would then, hypothetically, always lag behind. This should be taken into account too.

5. **To what extent is it important to independently search for additional information outside the model?**
   You cannot capture everything (like/dislike) within the model theoretically, it would not be smart to rely completely on such models. Especially within our domain it is paramount to search for information independently outside of the model before making a decision. Moreover, the performance of decisions is complex to measure as self-fulfilling prophecy forms a bottle neck. Say we refuse to admit a patient to the IC due to its extremely poor health conditions; as a consequence the patient will likely die (as how we predicted). But we do not know if the patient would survive in a hypothetical opposite situation in which we decide to admit him/her to the IC. The opposite of this example is also true and this always raises the question on the performance of our decisions. Nevertheless it is important to critically evaluate each and every decision, with or without the model.

Appendix B - Survey validation process

Within this appendix the survey validation process will be explicated. To perform a factor analysis, principal component analysis (PCA) is performed to translate a set of correlated variables into a reduced set of factors that capture the data (Fricker, Kulzy, & Appleget, 2012). By doing so, we aim to explore the extent to which the statements within the data actually correspond to the underlying theories for each of the factors. The PCA does not result in significant interpretations for each of the components, but can be used for exploratory purposes. In section 8.3 the initial PCA is explicated and substantiated with information regarding reliability of the data. Thereafter, in section 8.3 the final selection of components and meanings are argued. Finally, in section 8.3 a comparison is made between the initial constructs and the final selection and definition of components.

Initial PCA & reliability

With the test survey results from the relevant sample of decision-makers, dimension reduction was conducted using PCA. This method enables to test the hypothesized constructs within the survey. All statements within the survey were formulated and presented on the same Likert Scale. As the survey contains six hypothesized constructs, the analysis is conducted with a fixed number of six factors that should be extracted. With a varimax rotation method it was possible to rotate the component matrix and obtain sensible information.

The suitability of the data structure is tested by means of the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) and the Bartlet's Test of Sphericity. Both measures indicated the structure of the dataset could not be verified. This is due to a non-positive definitive correlation matrix. The cause for this hurdle is the presence of negative eigenvalues in the correlation matrix or linear dependencies among the variables. The main cause for both possible instabilities is the limited sample size of this study. However, a PCA can be conducted to capture the data into a reduced set of factor.

On that account, the rotated component matrix fig. 3 is analyzed to understand how the statements within the survey loads to each of the components. Although all factor loadings were allowed in the analysis, only those greater than 0.5 were perceived satisfactory to be considered in the final component structure. The first eight statements loaded to component 1 were analyzed with Cronbach's alpha, yielding $\alpha = 0.761$. The following four statements loaded on component 2 yielded $\alpha = 0.83$. Subsequently, the following four statements loaded on component 3 yielded $\alpha = 0.835$. The following two statements loaded on component 4 yielded $\alpha = 0.789$. The following three statements loaded on component 5 yielded $\alpha = 0.775$. The last three statements loaded on component 6 yielded $\alpha = 0.583$.

From the initial analysis and reliability test the statements can be selected and considered for an adjusted version of the HMA survey. The statements are selected based on the factor loadings and the initial theoretical construct it represents. By doing so, we are able to obtain suitable statements and form reliable components.

Figure 3: Initial PCA

Rotated Component Matrix[a]

	Component							
	1	2	3	4	5	6	Construct	Original construct
21 - The Councyl model enables me to critically reflect on my choices	0,833	0,393		0,183		-0,141	Construct A	CRIT
8 - Information about the development of the advice is clearly presented	0,819	0,194	0,150		-0,336	-0,231		Accessibility
22 - When I use the Councyl model, I can still critically consider my choices	0,818	0,214	0,222					CRIT
16 - The Councyl model stimulates the exchange of ideas within my team for making choices	0,723	0,117	-0,209	-0,451		0,113	Undecided	Get expertise
14 - Using the Councyl model encourages me less to approach my colleagues for additional advice	0,679	-0,129	0,156	-0,530	0,193			Get expertise
12 - The Councyl model enables me to trace the formation of the advice	-0,625	0,477	0,403		0,432			Tractability
11 - The Councyl model offers me the opportunity to trace back the advice on the basis of criteria and weights	0,569	0,483	0,421	-0,385				Tractability
17 - The Councyl model leads to an improvement of competences within my team	0,462	0,383	0,163	0,298	-0,179	0,202		Identify expertise
4 - Using the Councyl model does not diminish the moral responsibility I bear for my choices	0,252	0,928			0,105		Construct B	HMA
10 - The Councyl model enables me to sufficiently analyze the advice		0,924		-0,198	0,218	-0,122	Undecided	Tractability
1 - If I use the Councyl model, I can still take responsibility for my choices	0,147	0,815	0,195	0,343	-0,117		Construct B	HMA
18 - Using the Councyl model makes it more difficult to recognize (new) experts in my department	0,426	0,617	0,143	-0,152	0,316	0,445	Undecided	Identify expertise
20 - The Councyl model does not stand in the way of the development of competences of colleagues of mine and my colleagues	0,188		0,926	0,251			Construct C	Identify expertise
2 - When I use the Councyl model, I feel limited in making my choice	0,184	0,245	0,878	0,188	0,187	-0,137		HMA
5 - The Councyl model offers a sufficient degree of accessibility to relevant information	-0,450	-0,158	0,787			0,334		Accessibility
13 - The Councyl model offers me the opportunity to consult colleagues when making choices	-0,372	-0,101	-0,734		0,388	-0,153	Undecided	Get expertise
9 - The Councyl model enables me to study the development of the advice properly	-0,173	-0,198	0,126	0,862	0,238		Construct D	Tractability
3 - I consider the advice of the Councyl model to be non-binding	0,195	0,261	0,265	0,836	0,192			HMA
24 - The Councyl model provides sufficient information to be able to critically assess the advice		0,103			0,901		Construct E	CRIT
6 - The Councyl model allows me to understand how the system works	-0,217	0,233	-0,178		0,823	-0,215		Accessibility
23 - The Councyl model provides the appropriate information to be able to critically execute decision-making			0,238	0,234	0,701	0,188		CRIT
7 - The Councyl model offers the correct and relevant information for making choices	-0,156	-0,227		-0,425		0,835	Construct F	Accessibility
15 - The Councyl model leads to an improvement of the dialogue betweenme and my colleagues			0,208	0,431	-0,128	0,769		Get expertise
19 - The use of the Councyl model leads to an improvement in my professional knowledge	0,425		0,349	0,186	-0,205	-0,528	Undecided	Identify expertise

Extraction Method: Principal Component Analysis.

a. Rotation converged in 13 iterations.

Final PCA & reliability

With the selected question from the previous section, another dimension reduction is conducted with PCA. The extraction is set on a fixed number of six factors, by which 91.56% of the variance is explained. The PCA is conducted with a varimax rotation, as well as all other settings in the initial PCA. Similarly to the initial PCA, the KMO and Barlett's Test of Sphericity measures could not be computed due to a non-positive definitive correlation matrix. However, all selected statements appeared to load in a similar manner to all related components as in the initial PCA. On that account, the reliability of each component is tested with the computation of Cronbach's alpha.

On that account, the rotated component matrix (see fig. 4) is analyzed to understand how the statements within the survey loads to each of the components. Although all factor loadings were allowed in the analysis, only those greater than 0.5 were perceived satisfactory to be considered in the final component structure. The first three statements loaded to component 1 were analyzed with Cronbach's alpha, yielding $\alpha = 0.874$. The following two statements loaded on component 2 yielded $\alpha = 0.587$. Subsequently, the following three statements loaded on component 3 yielded $\alpha = 0.775$. The following two statements loaded on component 4 yielded $\alpha = 0.904$. The following two statements loaded on component 5 yielded $\alpha = 0.789$. The last two statements loaded on component 6 yielded $\alpha = 0.583$. The literature suggests an arbitrary rule for Cronbach's alpha, holding 0.7 to be a good rule of thumb to determine the internal consistency. However, it is decided to maintain the components within the analysis that resulted in a slightly lower Cronbach's alpha (component 2 and 6). Hence, an interpretation of those components have implications for the conclusion of this study.

Comparison of Hypothesized Constructs & Actual Components

Based on the final PCA and selection of the statements the final survey can be constructed. Within the first component, statement 21, 8 and 22 are selected. Given the formulation of the statements and the initial constructs for each of the statements, it is decided to rephrase this construct as: "Critically reflect the presented information. The second component includes statements 20, 2 and 5. Those components originate from three different constructs. We rephrase this component as "Feeling restricted by the technology". The third statement includes statement 24, 6 and 23, derived from two distinctive constructs initially. Those are now clustered as "Critical understanding the technology". The fourth component includes statements 1 and 4, both originate from the HMA construct. Hence, we decide to maintain this construct by means of those statements. The fifth component includes statements 9 and 3, which is rephrased as "Understand the technology and my role". The last component includes statement 7 and 15, being rephrased as "Improved dialogue on decision-making". Although the initial structure of the survey could not be maintained with the current validation, we are able to reorganize the survey to preserve it's practicality.

Figure 4: Final PCA

Rotated Component Matrix[a]

	Component						Construct	Original construct
	1	2	3	4	5	6		
8 - Information about the development of the advice is clearly presented	0,925		-0,274			-0,160	Construct A	Accessibility
22 - When I use the Councyl model, I can still critically consider my choices	0,844	0,210		0,249				CRIT
21 - The Councyl model enables me to critically reflect on my choices	0,836			0,311	0,133			CRIT
4 - Using the Councyl model does not diminish the moral responsibility I bear for my choices	0,304		0,179	0,917			Construct B	HMA
1 - If I use the Councyl model, I can still take responsibility for my choices	0,286	0,220		0,859	0,222			HMA
20 - The Councyl model does not stand in the way of the development of competences of colleagues of mine and my colleagues	0,321	0,904			0,180		Construct C	Identify expertise
2 - When I use the Councyl model, I feel limited in making my choice	0,236	0,892	0,116	0,270	0,180			HMA
5 - The Councyl model offers a sufficient degree of accessibility to relevant information	-0,457	0,794	-0,143	-0,101		0,339		Accessibility
9 - The Councyl model enables me to study the development of the advice properly	-0,130	0,134	0,193	-0,108	0,929		Construct D	Tractability
3 - I consider the advice of the Councyl model to be non-binding	0,252	0,295	0,134	0,367	0,812			HMA
6 - The Councyl model allows me to understand how the system works	-0,205	-0,136	0,889	0,139	0,101	-0,205	Construct E	Accessibility
24 - The Councyl model provides sufficient information to be able to critically assess the advice			0,821	0,122		-0,146		CRIT
23 - The Councyl model provides the appropriate information to be able to critically execute decision-making		0,239	0,793	-0,226	0,159	0,284		CRIT
15 - The Councyl model leads to an improvement of the dialogue betweenme and my colleagues		0,171	-0,148	0,111	0,314	0,877	Construct F	Get expertise
7 - The Councyl model offers the correct and relevant information for making choices	-0,263			-0,250	-0,473	0,774		Accessibility

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

Appendix C - Results descriptive analysis

This chapter presents the other results of the data analysis, as part of the descriptive analysis as presented in chapter 5. In section 8.3 the visualisations of the high-level analysis are included. Lastly, section 8.3 contains the distribution of scores for each of the theoretical constructs.

High-level analysis

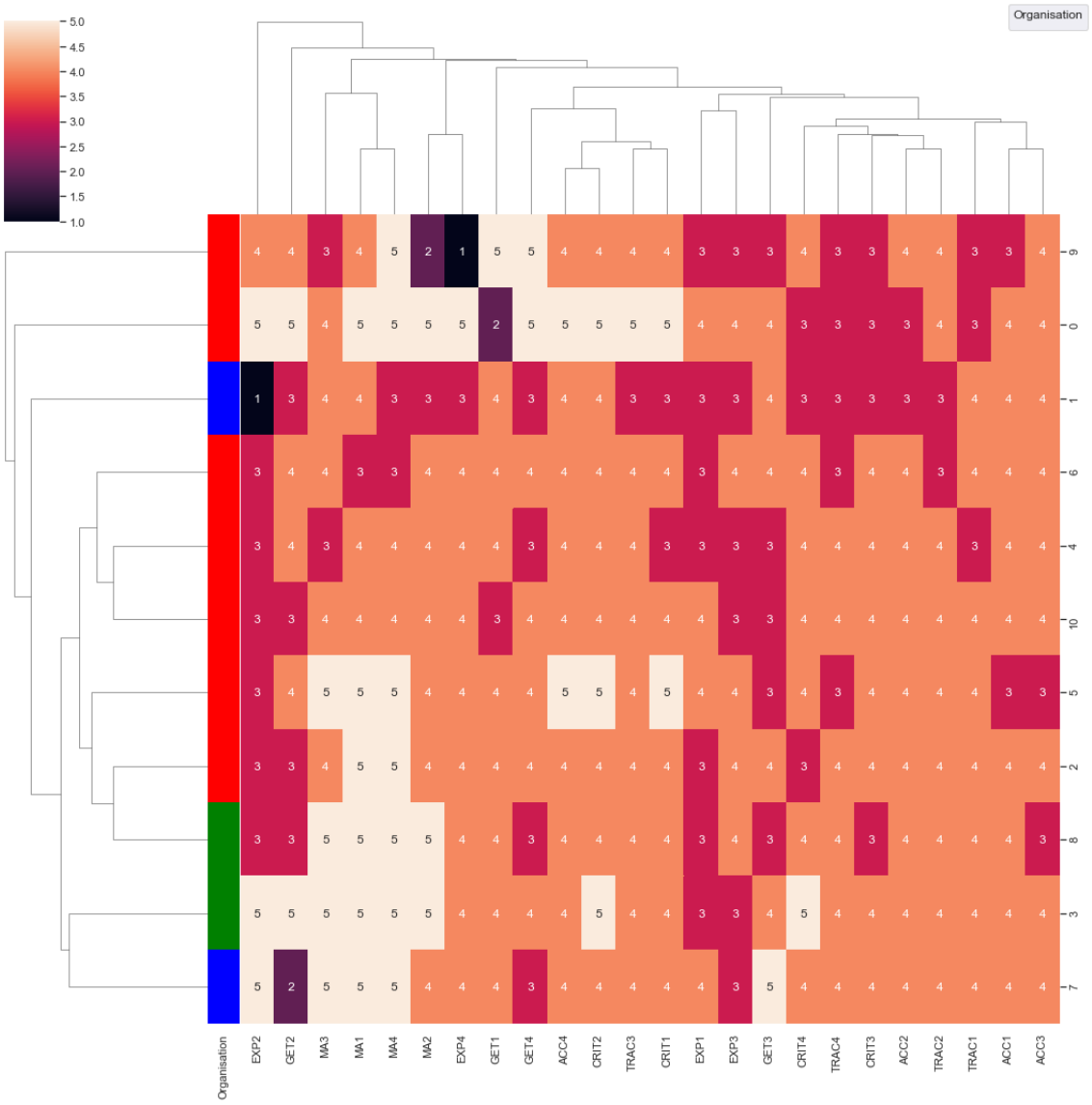


Figure 5: Dendrogram of heatmap, clustered based on organisation. Red = UMCG, Green = Deloitte and Blue = OLVG

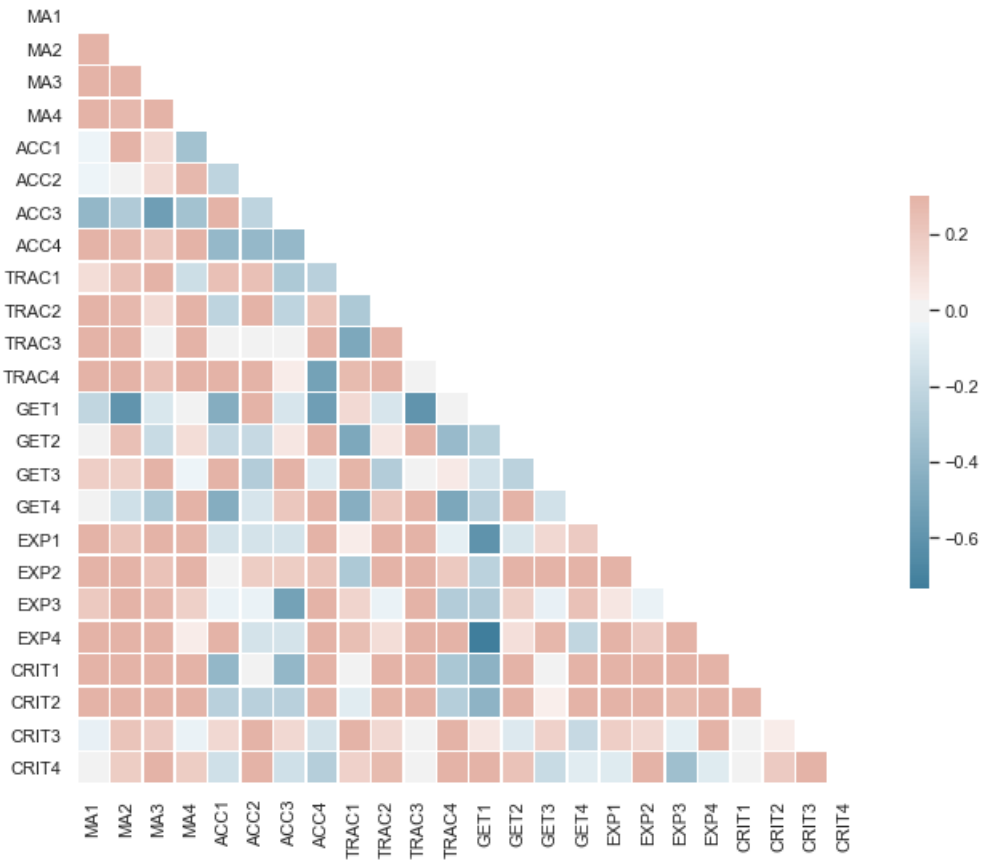


Figure 6: Correlation matrix all statement

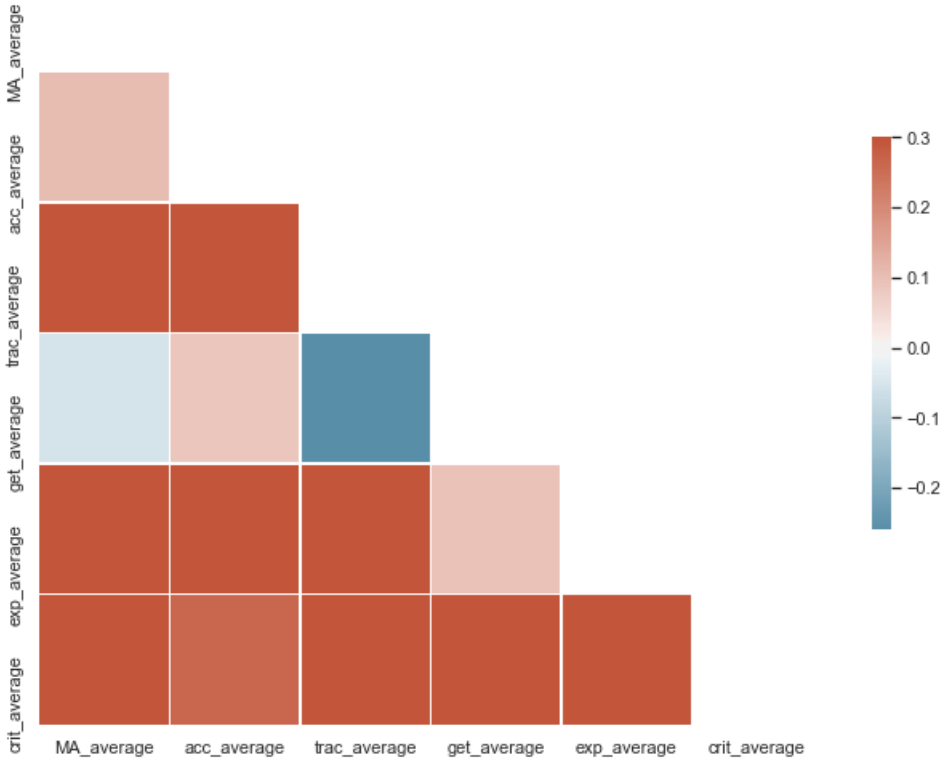


Figure 7: Correlation matrix all constructs

Distribution of scores

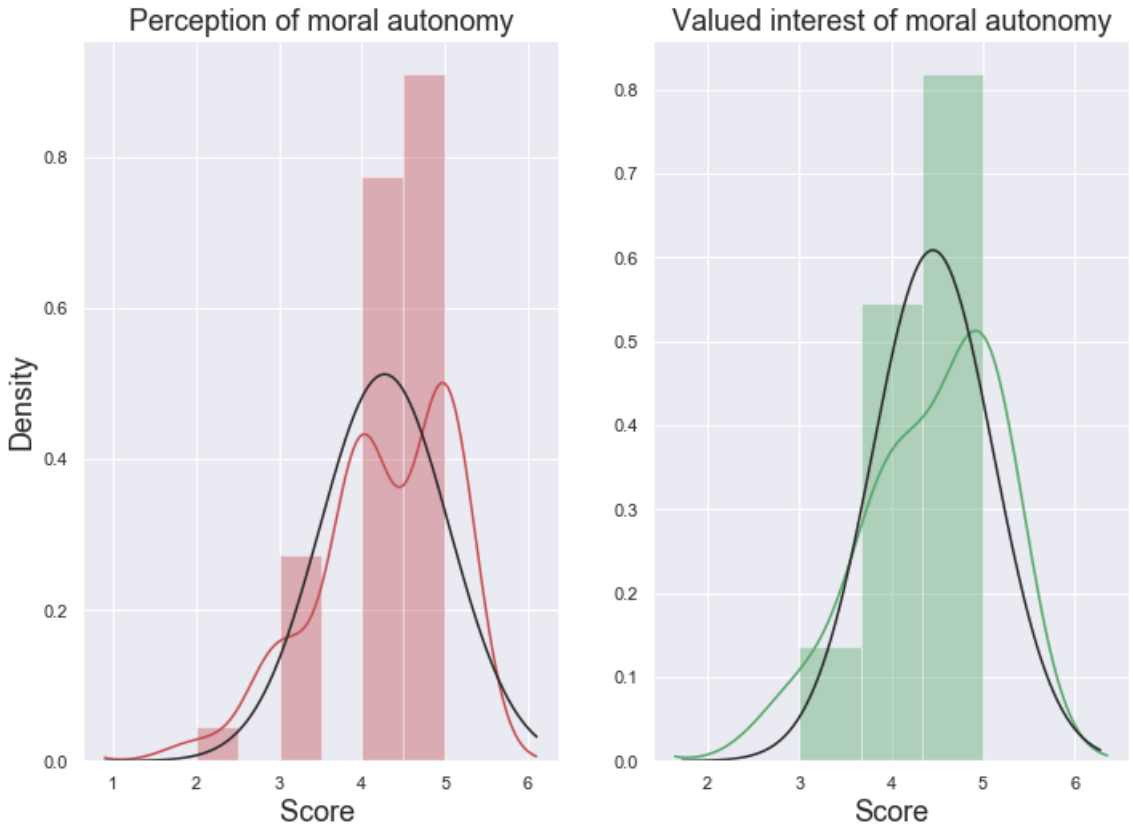


Figure 8: Distribution of scores HMA

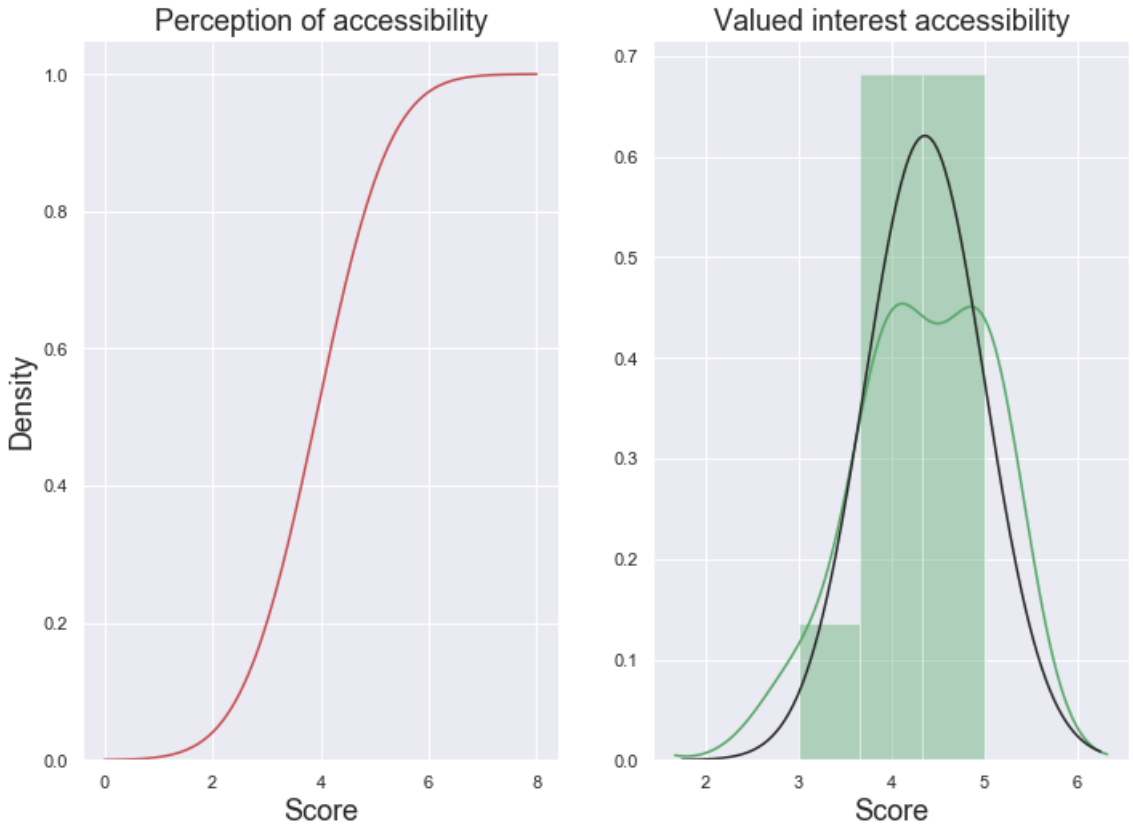


Figure 9: Distribution of scores accessibility

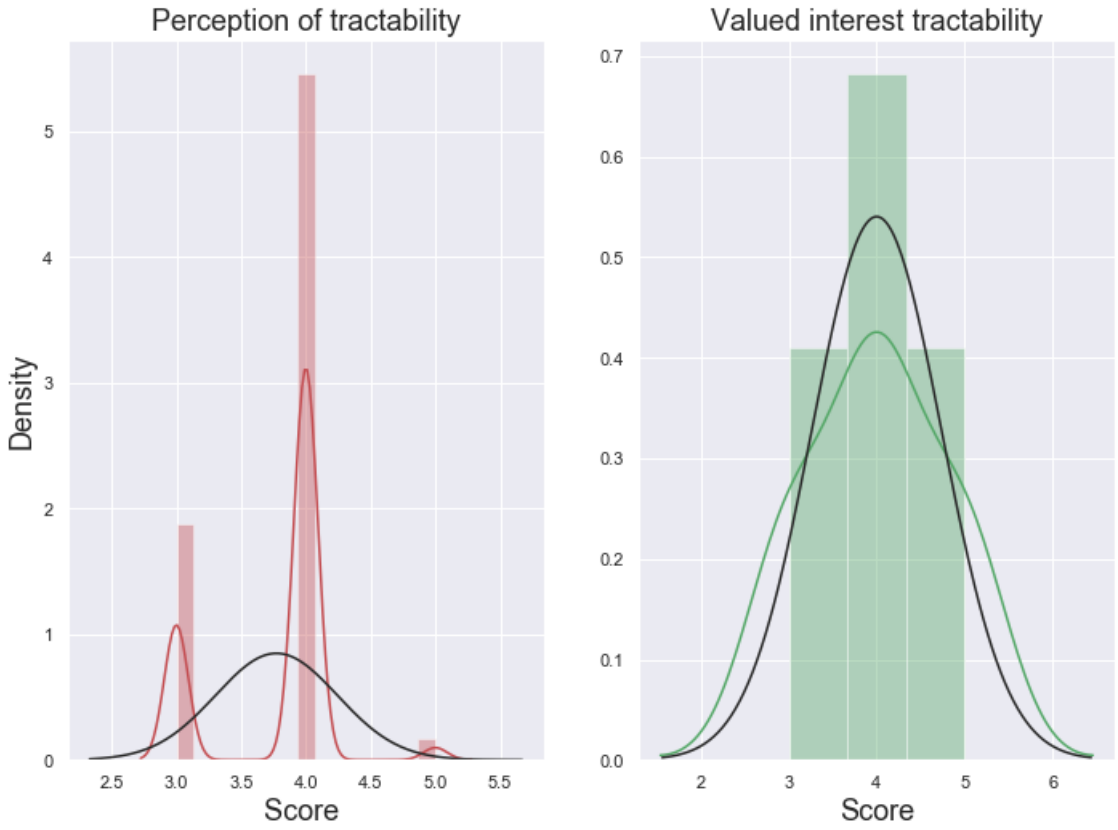


Figure 10: Distribution of scores tractability

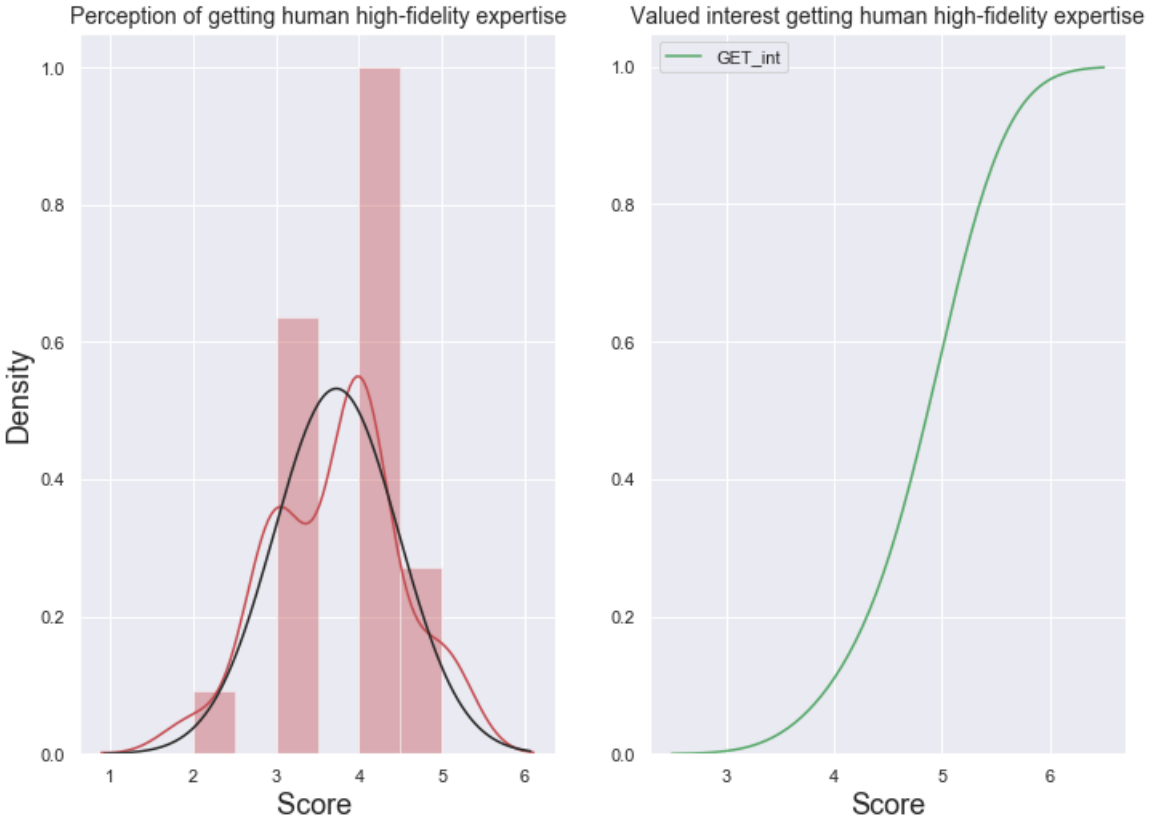


Figure 11: Distribution of scores getting human high-fidelity expertise

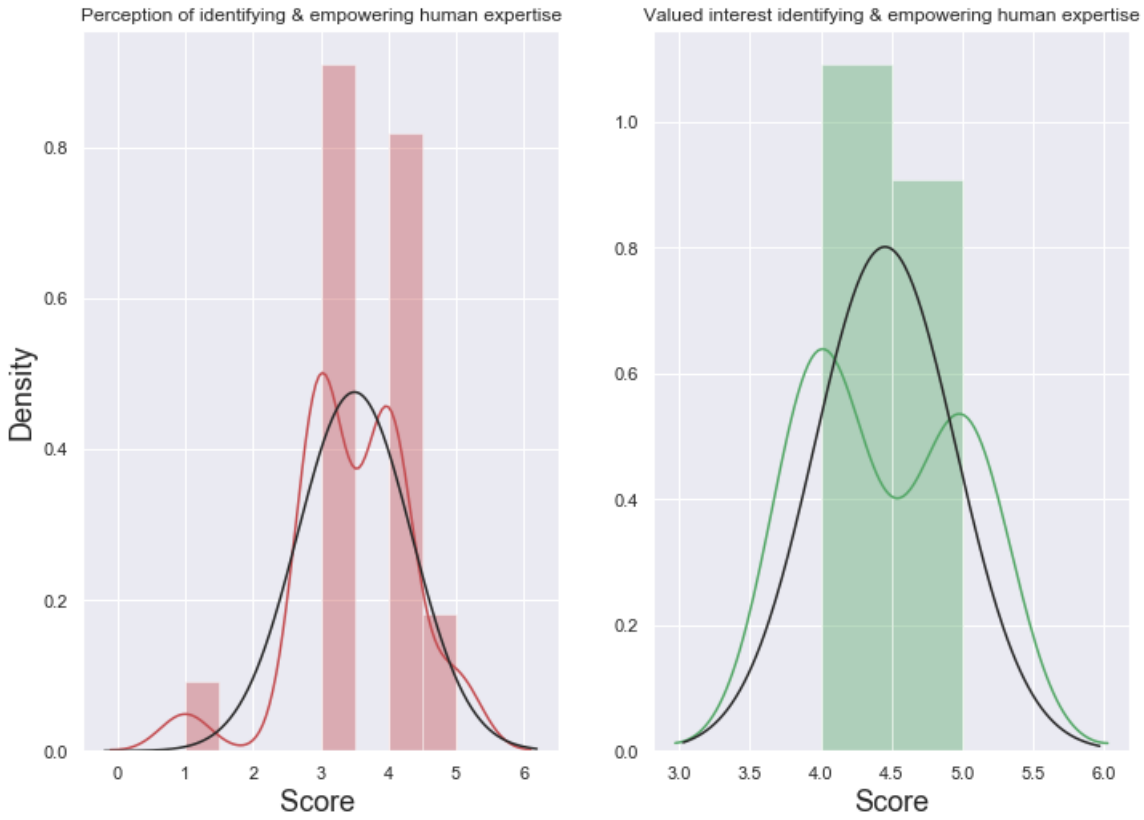


Figure 12: Distribution of scores identify & empower human expertise

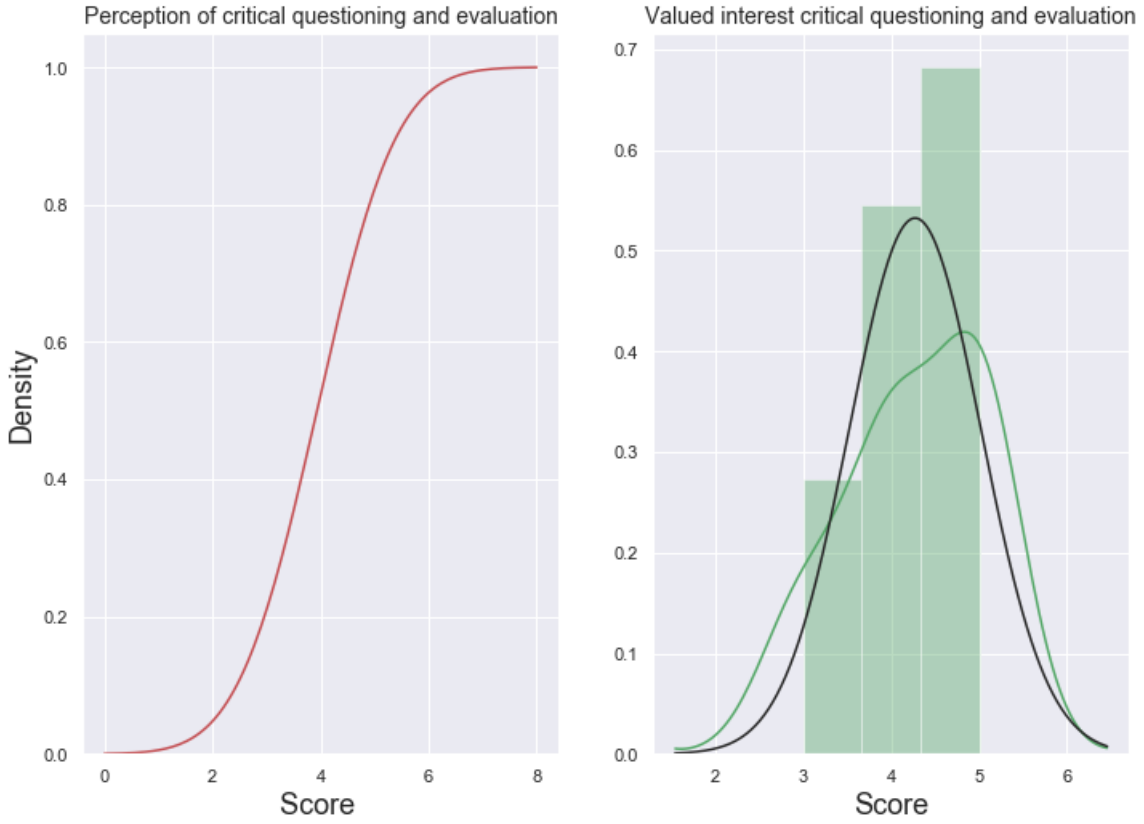


Figure 13: Distribution of scores critical questioning and evaluation

Significance test mean difference

This subsection entails the presentation of the t-test. We computed this test to examine the significance in the difference between the mean values of the perception and importance statements. This is done for each of the construct.

Construct	**Test statistic**	**p-value**
HMA	-0.6480335690083339	0.5243338446980463
Accessibility	-2.197934911319288	0.039896087619281935
Tractability	-0.922138891954146	0.3674471935283051
Get high-fidelity human expertise	-6.669240172520744	1.7150015265395292e-06
Identify & Empower human expertise	-4.183300132670376	0.0004582668669996776
Critical questioning and evaluation	-1.3105560849915565	0.2048590199651411

Relationship between constructs

Relationships on perception statements

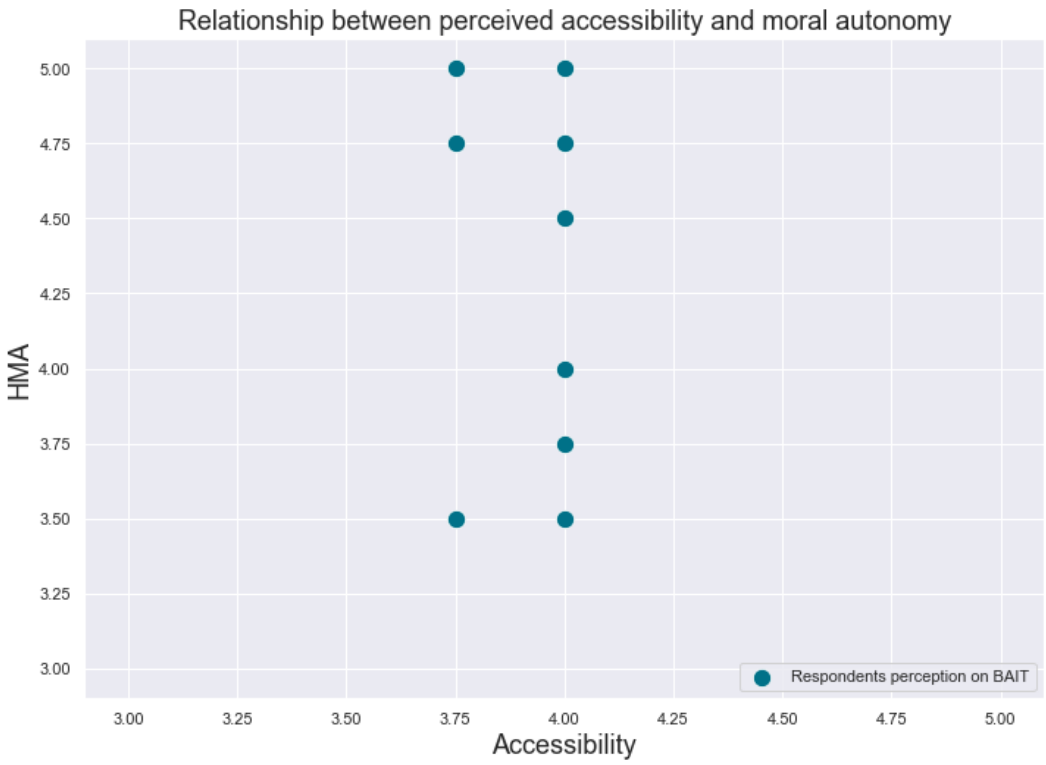


Figure 14: Perception of accessibility and HMA

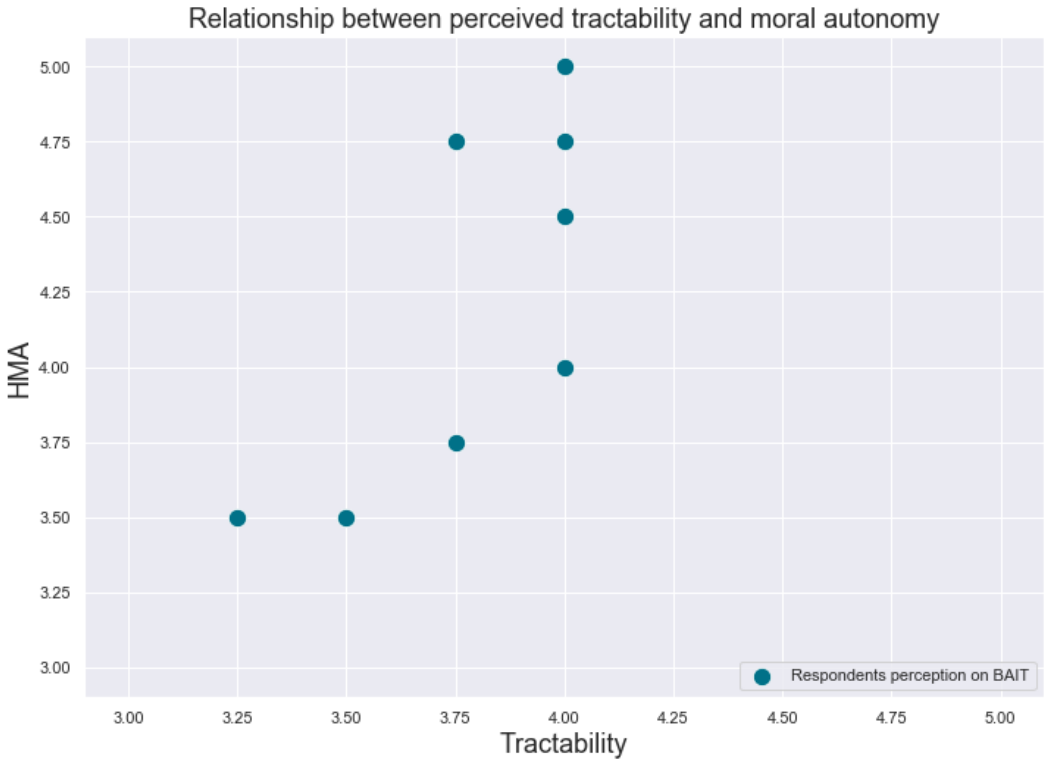


Figure 15: Perception of tractability and HMA

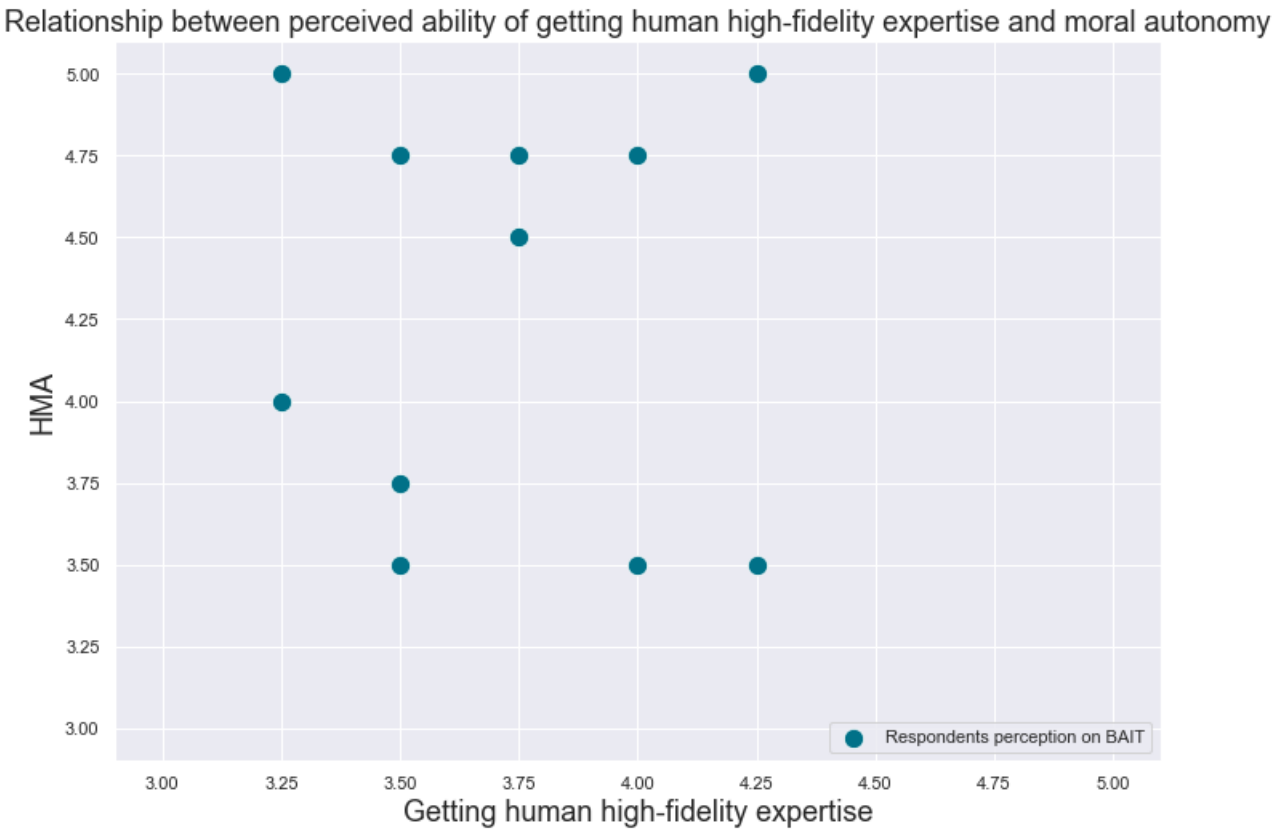


Figure 16: Perception of getting human expertise and HMA

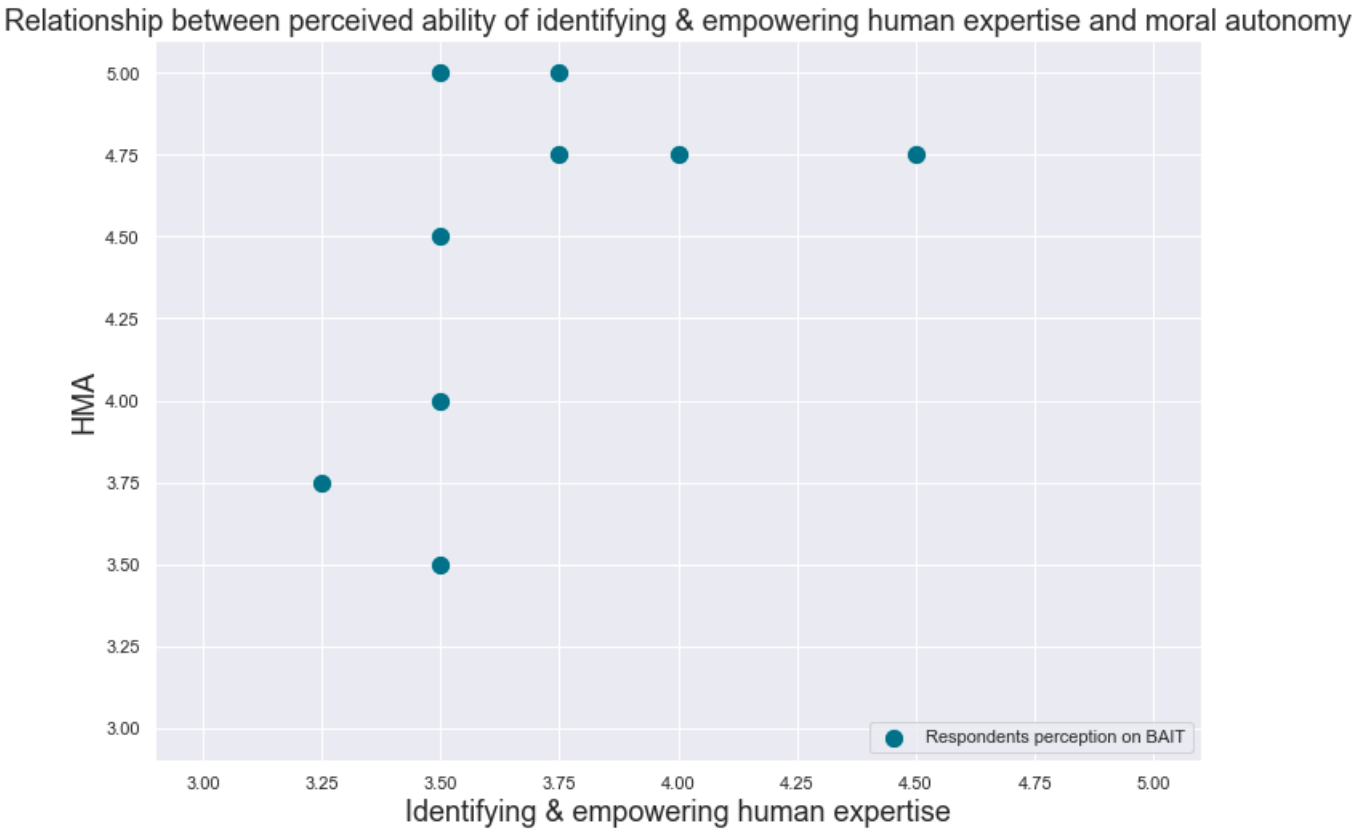


Figure 17: Perception of identifying human expertise and HMA

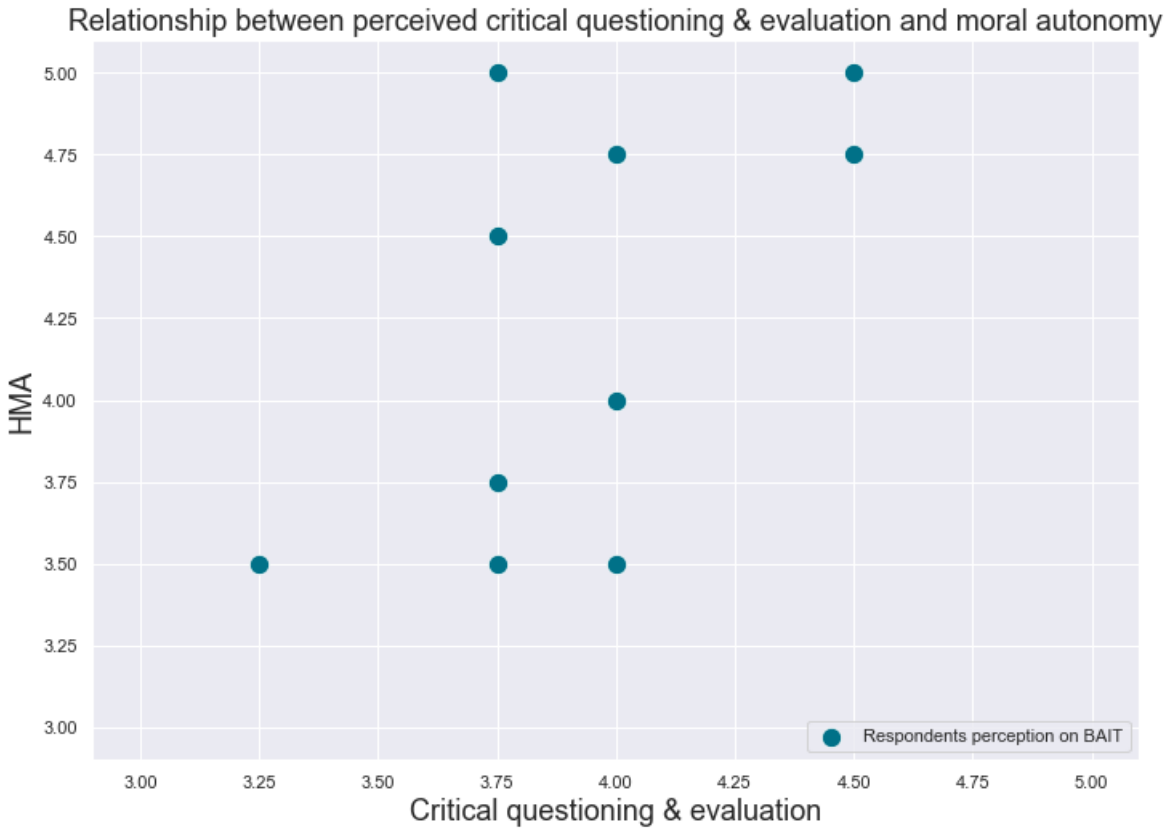


Figure 18: Perception of critical questioning and HMA

Relationships on importance statements

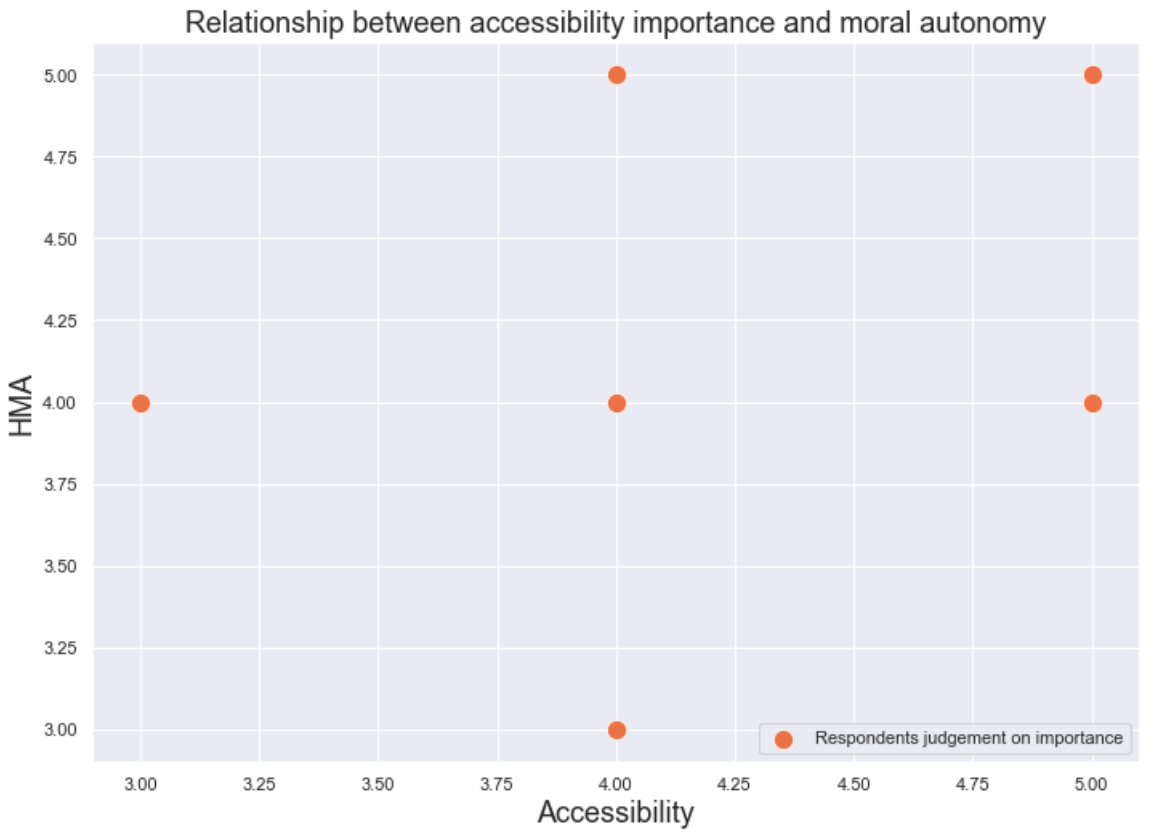


Figure 19: Importance of accessibility and HMA

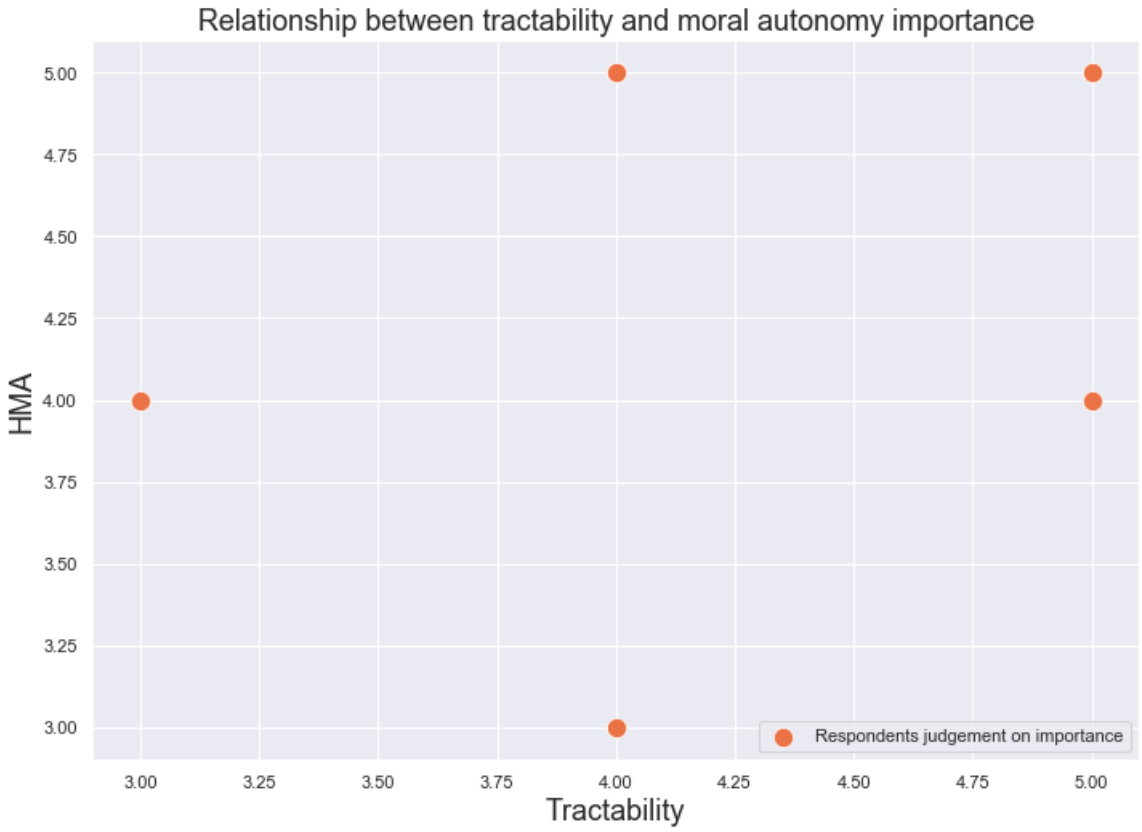


Figure 20: Importance of tractability and HMA

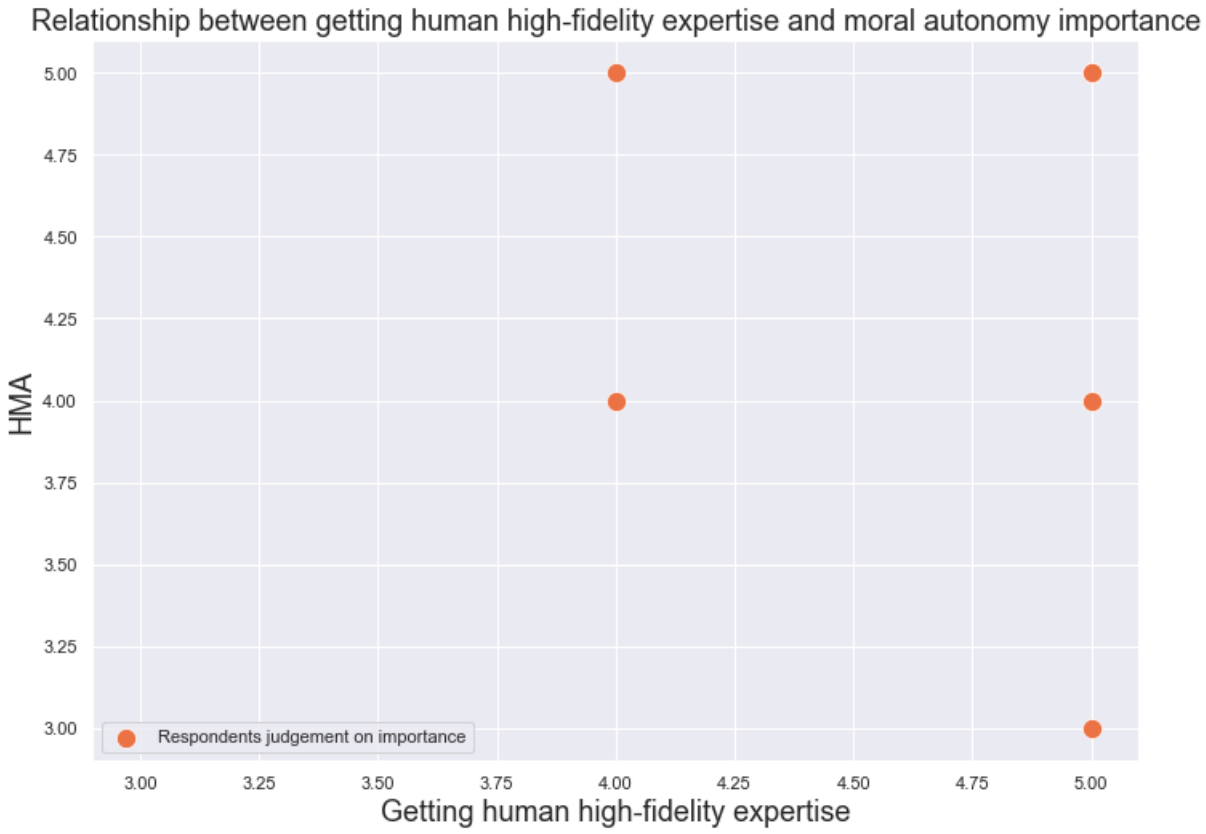


Figure 21: Importance of getting human expertise and HMA

Figure 22: Importance of identifying human expertise and HMA

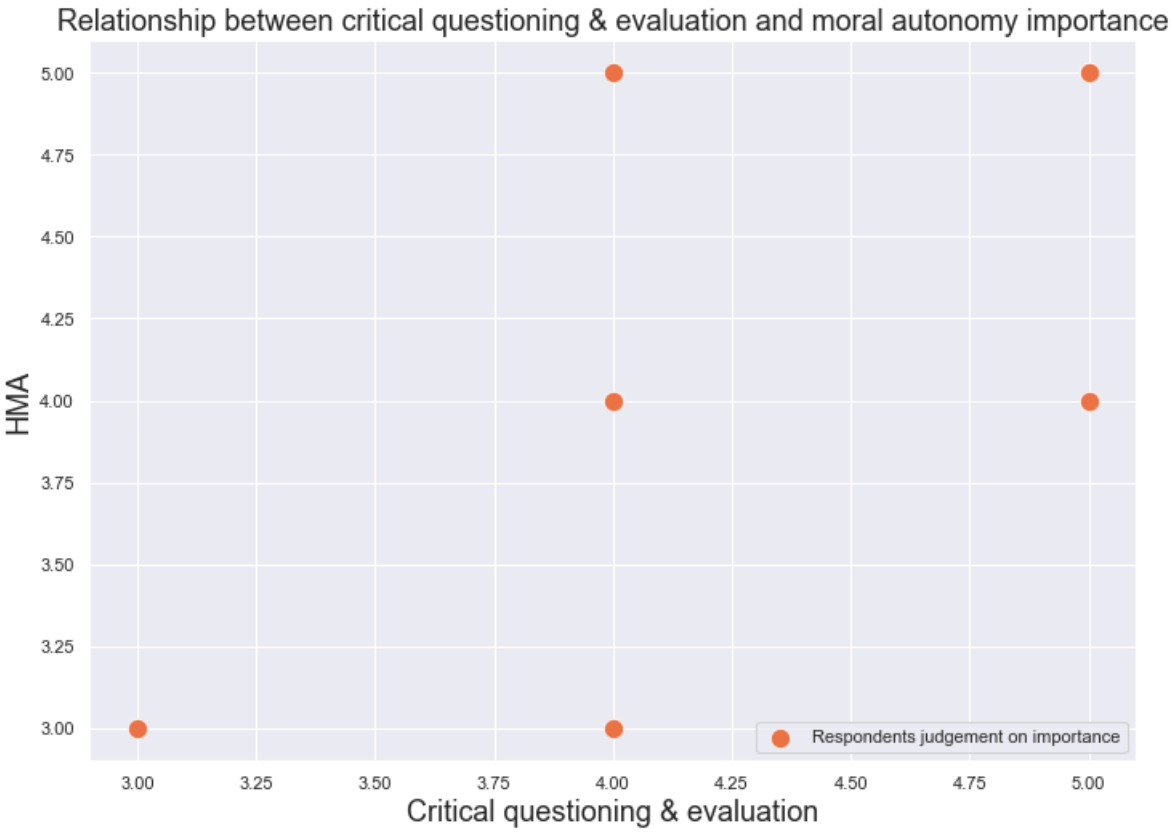


Figure 23: Importance of critical questioning and HMA

Appendix D - Questionnaire

This chapter shows the questionnaire as it is presented to the respondents. The English translations of the statements can be found in chapter 4.

Introductie

Introductie.

Bedankt dat u wil deelnemen aan deze vragenlijst over het Councyl model.

Voor de verdere ontwikkeling van onze technologie zijn wij benieuwd naar uw indrukken van het Councyl model.

Uw deelname aan dit onderzoek is geheel vrijwillig en u kunt zich op elk moment terugtrekken.
De enquête duurt maximaal 8 minuten. Uw antwoorden worden geanonimiseerd.
Vragen die zijn gemarkeerd met een sterretje (*), zijn vereist.
Als u vragen hebt over de enquête, kunt u mij een e-mail sturen op: can@councyl.ai
We stellen uw inbreng zeer op prijs.

Stellingen

Q1.
Wilt u voor de volgende stellingen aangeven of u het er mee eens bent of niet?

	Sterk mee oneens	Mee oneens	Neutraal	Mee eens	Sterk mee eens
Ik vind het van belang om toegang te hebben tot alle vormen van informatie die het Councyl model biedt	○	○	○	○	○
Ik vind het van belang dat ik inzicht heb in het gewicht van elke factor die in het Councyl model is opgenomen	○	○	○	○	○

	Sterk mee oneens	Mee oneens	Neutraal	Mee eens	Sterk mee eens
Ik vind het belangrijk dat ik de totstandkoming van het advies van het Councyl model kritisch en onafhankelijk kan beoordelen	○	○	○	○	○
Ik vind het belangrijk om niet alleen afhankelijk te zijn van het Councyl model bij het maken van keuzes	○	○	○	○	○
Ik vind het belangrijk dat het Councyl model de ontwikkeling van deskundigheid bij mij en mijn collega's stimuleert	○	○	○	○	○
Ik vind het belangrijk om altijd de optie te hebben tot het raadplegen van collega's binnen mijn afdeling bij het maken van keuzes	○	○	○	○	○

Q2. Bedankt voor uw inbreng tot dusver.

In het volgende onderdeel zijn stellingen geformuleerd over uw ervaringen met het Councyl model. Hiervoor kunt u de demo's uitproberen om een indruk te krijgen van het model. Nadat u één van de demo's bezocht heeft voor een impressie, kunt u in het volgende onderdeel de stellingen beantwoorden.

(Gelieve met rechtermuisknop te drukken op een link en de pagina te openen in een nieuw tabblad)

HR-demo
Vergunningen-demo
Verzekeringsaanvraag-demo

Q3. Wilt u voor de volgende stellingen aangeven of u het er mee eens bent of niet?

	Sterk mee oneens	Mee oneens	Neutraal	Mee eens	Sterk mee eens
Het Councyl model stimuleert de gedachtewisseling binnen mijn team voor het maken van keuzes	○	○	○	○	○
Als ik het Councyl model gebruik, kan ik nog steeds verantwoordelijkheid dragen voor mijn keuzes	○	○	○	○	○
Het Councyl model biedt voldoende informatie om het advies kritisch te kunnen beoordelen	○	○	○	○	○
Het gebruik van het Councyl model doet niets af aan de morele verantwoordelijkheid die ik draag voor mijn keuzes	○	○	○	○	○
Het Councyl model biedt de juiste en relevante informatie aan voor het maken van keuzes	○	○	○	○	○
Door gebruik van het Councyl model wordt het lastiger om op mijn afdeling (nieuwe) deskundigen te herkennen	○	○	○	○	○
Het Councyl model staat de ontwikkeling van competenties van collega's van mij en mijn collega's niet in de weg	○	○	○	○	○
Het Councyl model biedt een voldoende mate van toegankelijkheid tot relevante informatie	○	○	○	○	○
Het gebruik van het Councyl model leidt tot een verbetering van mijn vakkennis	○	○	○	○	○

	Sterk mee oneens	Mee oneens	Neutraal	Mee eens	Sterk mee eens
Als ik het Councyl model gebruik, voel ik mij beperkt in het maken van mijn keuze	○	○	○	○	○
Het Councyl model stelt mij in staat de totstandkoming van het advies goed te bestuderen	○	○	○	○	○
Het Councyl model stelt mij in staat de totstandkoming van het advies te traceren	○	○	○	○	○
Het advies van het Councyl model beschouw ik als vrijblijvend	○	○	○	○	○
Het gebruik van het Councyl model moedigt mij minder aan om mijn collega's te benaderen voor extra advies	○	○	○	○	○
Informatie over de totstandkoming van het advies wordt duidelijk gepresenteerd	○	○	○	○	○
Het Councyl model biedt mij de mogelijkheid om collega's te raadplegen bij het maken van keuzes	○	○	○	○	○
Als ik het Councyl model gebruik, kan ik nog steeds mijn keuzes kritisch afwegen	○	○	○	○	○
Het Councyl model leidt tot een verbetering van de dialoog tussen mij en mijn collega's	○	○	○	○	○
Het Councyl model verschaft de geschikte informatie om de besluitvorming kritisch te kunnen uitvoeren	○	○	○	○	○

	Sterk mee oneens	Mee oneens	Neutraal	Mee eens	Sterk mee eens
Het Councyl model stelt mij in staat het advies in voldoende mate te analyseren	○	○	○	○	○
Het Councyl model stelt mij in staat om te begrijpen hoe het systeem werkt	○	○	○	○	○
Het Councyl model leidt tot een verbetering van competenties binnen mijn team	○	○	○	○	○
Het Councyl model stelt mij in staat kritisch te reflecteren op mijn keuzes	○	○	○	○	○
Het Councyl model biedt mij de mogelijkheid het advies te herleiden op basis van criteria en gewichten	○	○	○	○	○

Persoonlijke informatie

Q1. Bij welke organisatie bent u werkzaam?

Deloitte

Onze Lieve Vrouwen Gasthuis (OLVG)

Universitair Medisch Centrum Groningen (UMCG)

Q2. Wat is uw functie binnen uw organisatie?

Neonatoloog

Chirurg

Anders, namelijk:

Q2. Wat is uw functie binnen uw organisatie?

Projectleider

Consultant

Anders, namelijk:

Q2. Wat is uw functie binnen uw organisatie?

Intensivist

Chirurg

Anders, namelijk:

Q3. Hoe lang bent u in dienst bij uw organisatie?

0 tot 5 jaar

5 tot 10 jaar

10 tot 15 jaar

Langer dan 15 jaar

Q4. Wat is uw hoogst genoten opleidingsniveau?

mbo

hbo

wo

gepromoveerd

Q5. Binnen welke leeftijdscategorie valt u?

18 tot 25 jaar

26 tot 35 jaar

36 tot 45 jaar

46 tot 55 jaar

56 tot 65 jaar

65+

Appendix E - Data analysis code

This chapter contains the data analysis code of the descriptive analysis. This chapter is included for reproducibility purposes, so similar codes can be used for each of the statements and constructs. Moreover, a quick analysis can be performed with the ready-to-use codes for producing visualisations.

Exploratory data analysis empirical study to human moral autonomy of BAIT

This notebook includes an exploratory data analysis on the moral autonomy of potential users (decision-makers) of Behavioural Artificial Intelligence Technology (BAIT). The analysis is step-wise conducted in which relationships were aimed to be identified. It mainly is subdivided into the following steps:

1. Import relevant libraries
2. Data pre-processing to make it ready
3. Generic visualisations to show personal characteristics and identify patterns
4. Distributions of scores per theoretical construct
5. Correlation graphs between constructs and theory

1. Import relevant libraries for data analysis

In [ ]:

```
import pandas as pd
import numpy as np
import os
import seaborn as sns
import matplotlib.pyplot as plt
sns.set(color_codes=True)
from scipy.stats import norm
```

Set working directory

In [ ]:

```
os.chdir('/Users/canyildiz/Downloads')
os.getcwd()
df = pd.read_csv('councyl2.csv')
```

2. Data pre-processing

In [ ]:

```
# Change column name organisation
df.rename(columns={'Q1': 'Organisation'},inplace=True)
df.rename(columns={'Q2': 'Jobtype_UMCG'},inplace=True)
df.rename(columns={'Q2.0': 'Jobtype_Deloitte'},inplace=True)
df.rename(columns={'Q2.1': 'Jobtype_OLVG'},inplace=True)


# Change all column names of theoretical constructs
df.rename(columns={'Q1_1':'MA_int', 'Q1_2':'ACC_int', 'Q1_3': 'TRAC_int', 'Q1_4':'GET_int', 'Q1_5':'EXP_int', 'Q1_6':
'CRIT_int','Q3_1': 'MA1', 'Q3_2': 'MA2', 'Q3_3': 'MA3',
       'Q3_4':'MA4', 'Q3_5':'ACC1', 'Q3_6':'ACC2', 'Q3_7':'ACC3', 'Q3_8':'ACC4', 'Q3_9':'TRAC1', 'Q3_10':'TRAC2', 'Q3_1
1':'TRAC3',
       'Q3_12':'TRAC4', 'Q3_13':'GET1', 'Q3_14':'GET2', 'Q3_15':'GET3', 'Q3_16':'GET4', 'Q3_17':'EXP1', 'Q3_18':'EXP2',
'Q3_19':'EXP3',
       'Q3_20':'EXP4', 'Q3_21':'CRIT1', 'Q3_22':'CRIT2', 'Q3_23':'CRIT3', 'Q3_24':'CRIT4'},inplace=True)

# Selection of theoretical constructs
perception_df = df.iloc[:,6:30]
other_df = df.iloc[:, 30:]
MA_df = perception_df.iloc[:, 0:4]
acc_df=perception_df.iloc[:, 4:8]
trac_df=perception_df.iloc[:, 8:12]
get_df=perception_df.iloc[:, 12:16]
exp_df=perception_df.iloc[:, 16:20]
crit_df=perception_df.iloc[:, 20:24]

# Taking the average per theoretical construct
MA_df_average = MA_df.mean(axis=1)
acc_df_average = acc_df.mean(axis=1)
trac_df_average = trac_df.mean(axis=1)
get_df_average = get_df.mean(axis=1)
exp_df_average = exp_df.mean(axis=1)
crit_df_average = crit_df.mean(axis=1)

# Adding averages of constructs to new columns
df['perceived_moral_autonomy'] = MA_df_average
df['perceived_accessibility'] = acc_df_average
df['perceived_tractability'] = trac_df_average
df['perceived_get_human_high_fidelity_expertise'] = get_df_average
df['perceived_identify_and_empower_human_expertise'] = exp_df_average
df['perceived_critical_questioning_and_evaluation'] = crit_df_average

# Concatenate questions per construct
ma_concat = pd.concat([df.MA1,df.MA2,df.MA3,df.MA4])
acc_concat = pd.concat([df.ACC1,df.ACC2,df.ACC3,df.ACC4])
trac_concat = pd.concat([df.TRAC1,df.TRAC2,df.TRAC3,df.TRAC4])
get_concat = pd.concat([df.GET1,df.GET2,df.GET3,df.GET4])
exp_concat = pd.concat([df.EXP1,df.EXP2,df.EXP3,df.EXP4])
crit_concat = pd.concat([df.CRIT1,df.CRIT2,df.CRIT3,df.CRIT4])

# Replace numeric values for organisation name
df['Organisation'] =df['Organisation'].replace(1,"Deloitte")
df['Organisation'] =df['Organisation'].replace(2,"OLVG")
df['Organisation'] =df['Organisation'].replace(3,"UMCG")

# Replace numeric values (strings) with job types
df['Jobtype_UMCG'] =df['Jobtype_UMCG'].replace(str(1),"Neonatoloog")
df['Jobtype_UMCG'] =df['Jobtype_UMCG'].replace(str(2),"Chirurg")
df['Jobtype_UMCG'] =df['Jobtype_UMCG'].replace(str(3),"Fellow neonatologie")
df['Jobtype_Deloitte'] =df['Jobtype_Deloitte'].replace(str(2),"Consultant")
df['Jobtype_Deloitte'] =df['Jobtype_Deloitte'].replace(str(3),"Eindverantwoordelijke")
df['Jobtype_OLVG'] =df['Jobtype_OLVG'].replace(str(1),"Intensivists")

# Create new table with average of perceptions
average_df['MA_average'] = pd.DataFrame(MA_df_average)
average_df['acc_average'] = acc_df_average
average_df['trac_average'] =trac_df_average
average_df['get_average'] =get_df_average
average_df['exp_average'] =exp_df_average
average_df['crit_average'] =crit_df_average
```

In [ ]:

```
average_df = pd.DataFrame(MA_df_average)
```

In [ ]:

```
average_df = average_df.iloc[:,1:]
average_df
```

In [ ]:

```
df.iloc[:,:6].mean().mean()
```

In [ ]:

```
df
```

3. Generic visualisations to show personal characteristics and identify patterns

In [ ]:

```
# Show number of respondents with barplot
f, ax = plt.subplots(figsize=(12, 8))
plot = sns.barplot(x=df.Organisation.value_counts().index, y=df.Organisation.value_counts(), label = True, palette="Set
1")
#plot.set(title = 'Score per respondent on each theoretical construct',fontsize=18)
ax.set(xlabel='Organisation', ylabel='Number of responses')
ax.set_title('Score per respondent on each theoretical construct',fontsize=18)
plt.show()

#plt.savefig('output_9_0.png')
```

In [ ]:

```
# Make figure and axes
fig, axs = plt.subplots(1, 3,figsize=(20,20))

# Show distribution of job types, UMCG
data_umcg = df.Jobtype_UMCG.value_counts().drop(" ")
labels_umcg = ["Chirurg", "Neonatoloog", "Fellow neonatologie"]
# Show distribution of job types, Deloitte
data_deloitte = df.Jobtype_Deloitte.value_counts().drop(" ")
labels_deloitte = ["Consultant", "Eindverantwoordelijke"]
# Show distribution of job types, OLVG
data_dolvg = df.Jobtype_OLVG.value_counts().drop(" ")
labels_olvg = ["Intensivists"]


# Creating plot UMCG
axs[0].pie(x = data_umcg,labels = labels_umcg, explode=[0.05]*3, shadow=True,autopct="%.1f%%")
# Creating plot Deloitte
axs[1].pie(x = data_deloitte,labels = labels_deloitte, explode=[0.05]*2, shadow=True,autopct="%.1f%%")
# Creating plot OLVG
axs[2].pie(x = data_dolvg,labels = labels_olvg, explode=[0.05]*1, shadow=True,autopct="%.1f%%")

# Set titles
axs[0].set_title('Job types UMCG', fontsize=18)
axs[1].set_title('Job types Deloitte', fontsize=18)
axs[2].set_title('Job types OLVG', fontsize=18)


# show plot
plt.show()
```

In [ ]:

```
# Draw a heatmap with the numeric values in each cell
f, ax = plt.subplots(figsize=(15, 10))
plot = sns.heatmap(perception_df, annot=True, fmt="d", linewidths=.5, ax=ax)
plot.set_title('Score per respondent on each statement', fontsize=18)
plt.xlabel("Statement number and type", fontsize=13)
plt.show()
```

In [ ]:

```
# Draw a clustermap/dendrogram to identify relationships between respondents based on organisation
# Organisation is of categorical type and hence can be used to identify those types of relations. As we can observe her
e
#https://stackoverflow.com/questions/62473426/display-legend-of-seaborn-clustermap-corresponding-to-the-row-colors
#A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an
 output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to c
lusters.

organisation_colors = df.Organisation.map({
    'UMCG':'red', 'Deloitte':'blue', 'OLVG':'green'
})

sns.clustermap(perception_df, annot = True, figsize= (15,15),row_colors = organisation_colors,metric="euclidean")
plt.legend(organisation_colors, title='Organisation',bbox_to_anchor=(1, 1), bbox_transform=plt.gcf().transFigure, loc=
'upper right')
plt.show()
```

In [ ]:

```
# https://stackoverflow.com/questions/42712304/seaborn-heatmap-subplots-keep-axis-ratio-consistent
#Draw correlation matrix between statements within a construct to show relation

# Compute the correlation matrix
corr1 = MA_df.corr()
corr2 = acc_df.corr()
corr3 = trac_df.corr()
corr4 = get_df.corr()
corr5 = exp_df.corr()
corr6 = crit_df.corr()

# Generate a mask for the upper triangle
mask1 = np.triu(np.ones_like(corr1, dtype=bool))
mask2 = np.triu(np.ones_like(corr2, dtype=bool))
mask3 = np.triu(np.ones_like(corr3, dtype=bool))
mask4 = np.triu(np.ones_like(corr4, dtype=bool))
mask5 = np.triu(np.ones_like(corr5, dtype=bool))
mask6 = np.triu(np.ones_like(corr6, dtype=bool))

# Set up the matplotlib figure
#f, ax = plt.subplots(figsize=(11, 9))
f,(ax1,ax2,ax3,ax4,ax5,ax6) = plt.subplots(1,6,sharey=False,figsize=(15, 11))
ax6.get_shared_y_axes().join(ax1,ax2,ax3,ax4,ax5)

# Generate a custom diverging colormap
cmap = sns.diverging_palette(230, 20, as_cmap=True)



# Draw the heatmap with the mask and correct aspect ratio
hm1 = sns.heatmap(corr1, mask=mask1, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar=False,ax=ax1)
hm2 = sns.heatmap(corr2, mask=mask2, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar=False,ax=ax2)
hm3 = sns.heatmap(corr3, mask=mask3, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar=False,ax=ax3)
hm4 = sns.heatmap(corr4, mask=mask4, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar=False,ax=ax4)
hm5 = sns.heatmap(corr5, mask=mask5, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar=False,ax=ax5)
hm6 = sns.heatmap(corr6, mask=mask6, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5},ax=ax6)
plt.show()
```

In [ ]:

```
# Draw correlation matrix across all statements to find more overall correlation between statements across constructs

# Compute the correlation matrix
corr = perception_df.corr()

# Generate a mask for the upper triangle
mask = np.triu(np.ones_like(corr, dtype=bool))

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(11, 9))

# Generate a custom diverging colormap
cmap = sns.diverging_palette(230, 20, as_cmap=True)

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5})
```

In [ ]:

```
# Correlation matrix between theoretical constructs as a whole

# Compute the correlation matrix
corr = average_df.corr()

# Generate a mask for the upper triangle
mask = np.triu(np.ones_like(corr, dtype=bool))

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(11, 9))

# Generate a custom diverging colormap
cmap = sns.diverging_palette(230, 20, as_cmap=True)

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5})
```

4. Distributions of scores per theoretical construct

In [ ]:

```
# Histogram. Distributions of scores, based on perception (red) and interest (green)
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=2,figsize=a4_dims)
sns.distplot(perception_df,ax=axs[0], bins = 5, kde=True, hist_kws=dict(ec="k"), fit = norm, color = 'r',label="Percept
ion scores")
sns.distplot(interest_df,ax=axs[1], bins = 5, kde=True, hist_kws=dict(ec="k"), fit = norm, color = 'g',label="Interest
 scores")

axs[0].set_title('Overall distribution of perception statements', fontsize=15)
axs[0].set_xlabel('Score ', fontsize=18)
axs[0].set_ylabel('Density', fontsize=18)
axs[1].set_title('Overall distribution of interest statements', fontsize=15)
axs[1].set_xlabel('Score', fontsize=18)
fig.legend()
plt.show()
```

In [ ]:

```
sns.distplot(perception_df, 10, kde=True, hist_kws=dict(ec="k"), fit = norm)
```

In [ ]:

```
sns.distplot(interest_df, 10, kde=True, hist_kws=dict(ec="k"), fit = norm)
```

In [ ]:

```
# Histogram. Distributions of scores, based on perception (red) and interest (green)
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=2,figsize=a4_dims)
sns.distplot(ma_concat,ax=axs[0],fit=norm, color = 'r')
sns.distplot(df.MA_int,ax=axs[1], fit=norm, color = 'g')

axs[0].set_title('Perception of moral autonomy', fontsize=18)
axs[0].set_xlabel('Score ', fontsize=18)
axs[0].set_ylabel('Density', fontsize=18)
axs[1].set_title('Valued interest of moral autonomy', fontsize=18)
axs[1].set_xlabel('Score', fontsize=18)

plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=2,figsize=a4_dims)
#sns.distplot(acc_concat,ax=axs[0],fit=norm, color = 'r')
sns.kdeplot(acc_concat,ax=axs[0], cumulative=True,bw=1, color = 'r', legend = True)
sns.distplot(df.ACC_int,ax=axs[1], fit=norm, color = 'g')

axs[0].set_title('Perception of accessibility', fontsize=18)
axs[0].set_xlabel('Score', fontsize=18)
axs[0].set_ylabel('Density', fontsize=18)
axs[1].set_title('Valued interest accessibility', fontsize=18)
axs[1].set_xlabel('Score', fontsize=18)

plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=2,figsize=a4_dims)
sns.distplot(trac_concat,ax=axs[0],fit=norm, color = 'r')
sns.distplot(df.TRAC_int,ax=axs[1], fit=norm, color = 'g')

axs[0].set_title('Perception of tractability', fontsize=18)
axs[0].set_xlabel('Score ', fontsize=18)
axs[0].set_ylabel('Density', fontsize=18)
axs[1].set_title('Valued interest tractability', fontsize=18)
axs[1].set_xlabel('Score', fontsize=18)

plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=2,figsize=a4_dims)
sns.distplot(get_concat,ax=axs[0],fit=norm, color = 'r')
#sns.distplot(interest_df.Q1_4,ax=axs[1], fit=norm, color = 'g', bw=1.5)
sns.kdeplot(df.GET_int,ax=axs[1], cumulative=True,bw=0.5, color = 'g')

axs[0].set_title('Perception of getting human high-fidelity expertise', fontsize=14)
axs[0].set_xlabel('Score ', fontsize=18)
axs[0].set_ylabel('Density', fontsize=18)
axs[1].set_title('Valued interest getting human high-fidelity expertise', fontsize=14)
axs[1].set_xlabel('Score', fontsize=18)

plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=2,figsize=a4_dims)
sns.distplot(exp_concat,ax=axs[0],fit=norm, color = 'r')
sns.distplot(df.EXP_int,ax=axs[1], fit=norm, color = 'g')

axs[0].set_title('Perception of identifying & empowering human expertise', fontsize=12)
axs[0].set_xlabel('Score ', fontsize=18)
axs[0].set_ylabel('Density', fontsize=18)
axs[1].set_title('Valued interest identifying & empowering human expertise', fontsize=12)
axs[1].set_xlabel('Score', fontsize=18)

plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=2,figsize=a4_dims)
#sns.distplot(crit_concat,ax=axs[0],fit=norm, color = 'r')
sns.kdeplot(crit_concat,ax=axs[0], cumulative=True,bw=1, color = 'r')
sns.distplot(df.CRIT_int,ax=axs[1], fit=norm, color = 'g')

axs[0].set_title('Perception of critical questioning and evaluation', fontsize=14)
axs[0].set_xlabel('Score ', fontsize=18)
axs[0].set_ylabel('Density', fontsize=18)
axs[1].set_title('Valued interest critical questioning and evaluation', fontsize=14)
axs[1].set_xlabel('Score', fontsize=18)

plt.show()
```

In [ ]:

```
perception_df
```

In [ ]:

```
sns.distplot(MA_df, bins=10, kde=False, hist_kws=dict(ec="k"))
```

In [ ]:

```
sns.distplot(interest_df, bins=10, kde=False, hist_kws=dict(ec="k"))
```

5. Correlation graphs between constructs and theory

In [ ]:

```
# Draw scatterplots with regression to make comparison with theory behind constructs. First perception plots (5x) and s
econd interest plots (5x)

a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=1,figsize=a4_dims)
plt.plot([0,1, 2, 3, 4,5], label='Assumed theory of Van den Hoven')
plot = sns.regplot(data=df, x="perceived_accessibility", y="perceived_moral_autonomy", truncate=False, color = "purple"
, label = 'Respondents perception on BAIT')
plot.set(xlim=(0, 5),ylim=(0,5))
plot.set_title('Relationship between perceived accessibility and moral autonomy', fontsize=18)
plt.legend(loc='lower right')
plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=1,figsize=a4_dims)
plt.plot([0,1, 2, 3, 4,5], label='Assumed theory of Van den Hoven')
plot = sns.regplot(data=df, x="perceived_tractability", y="perceived_moral_autonomy", truncate=False,color = "purple",
label = 'Respondents perception on BAIT')
plot.set(xlim=(0, 5),ylim=(0,5))
plot.set_title('Relationship between perceived tractability and moral autonomy', fontsize=18)
plt.legend(loc='lower right')
plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=1,figsize=a4_dims)
plt.plot([0,1, 2, 3, 4,5], label='Assumed theory of Van den Hoven')
plot = sns.regplot(data=df, x="perceived_get_human_high_fidelity_expertise", y="perceived_moral_autonomy",truncate=Fals
e,color = "purple", label = 'Respondents perception on BAIT')
plot.set(xlim=(0, 5),ylim=(0,5))
plot.set_title('Relationship between perceived ability of getting human high-fidelity expertise and moral autonomy', fo
ntsize=18)
plt.legend(loc='lower right')
plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=1,figsize=a4_dims)
plt.plot([0,1, 2, 3, 4,5], label='Assumed theory of Van den Hoven')
plot = sns.regplot(data=df, x="perceived_identify_and_empower_human_expertise", y="perceived_moral_autonomy", truncate=
False,color = "purple", label = 'Respondents perception on BAIT')
plot.set(xlim=(0, 5),ylim=(0,5))
plot.set_title('Relationship between perceived ability of identifying & empowering human expertise and moral autonomy',
fontsize=18)
plt.legend(loc='lower right')
plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=1,figsize=a4_dims)
plt.plot([0,1, 2, 3, 4,5], label='Assumed theory of Van den Hoven')
plot = sns.regplot(data=df, x="perceived_critical_questioning_and_evaluation", y="perceived_moral_autonomy", truncate=F
alse,color = "purple", label = 'Respondents perception on BAIT')
plot.set(xlim=(0, 5),ylim=(0,5))
plot.set_title('Relationship between perceived ability of critical questioning & evaluation and moral autonomy', fontsi
ze=18)
plt.legend(loc='lower right')
plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=1,figsize=a4_dims)
plt.plot([0,1, 2, 3, 4,5], label='Assumed theory of Van den Hoven')
plot = sns.regplot(data=df, x="ACC_int", y="MA_int", truncate=False,color = "green", label = 'Respondents judgement on
 importance')
plot.set(ylim=(0,5.2))
plot.set_title('Relationship between valued interest accessibility and moral autonomy', fontsize=18)
plt.legend(loc='lower right')
plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=1,figsize=a4_dims)
plt.plot([0,1, 2, 3, 4,5], label='Assumed theory of Van den Hoven')
plot = sns.regplot(data=df, x="TRAC_int", y="MA_int", truncate=False,color = "green", label = 'Respondents judgement on
importance')
plot.set(ylim=(0,5.2))
plot.set_title('Relationship between valued interest tractability and moral autonomy', fontsize=18)
plt.legend(loc='lower right')
plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=1,figsize=a4_dims)
plt.plot([0,1, 2, 3, 4,5], label='Assumed theory of Van den Hoven')
plot = sns.regplot(data=df, x="GET_int", y="MA_int", truncate=False,color = "green", label = 'Respondents')
plot.set(ylim=(0,5.2))
plot.set_title('Relationship between valued interest of getting human high-fidelity expertise and moral autonomy', font
size=18)
plt.legend(loc='lower right')
plt.savefig('1.png')
plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=1,figsize=a4_dims)
plt.plot([0,1, 2, 3, 4,5], label='Assumed theory of Van den Hoven')
plot = sns.regplot(data=df, x="EXP_int", y="MA_int", truncate=False,color = "green", label = 'Respondents')
plot.set(ylim=(0,5.2))
plot.set_title('Relationship between valued interest of identifying & empowering human expertise and moral autonomy', f
ontsize=18)
plt.legend(loc='lower right')
plt.savefig('2.png')
plt.show()
```

In [ ]:

```
a4_dims = (11.7, 8.27)
fig, axs = plt.subplots(ncols=1,figsize=a4_dims)
plt.plot([0,1, 2, 3, 4,5], label='Assumed theory of Van den Hoven')
plot = sns.regplot(data=df, x="CRIT_int", y="MA_int", truncate=False,color = "green", label = 'Respondents')
plot.set(ylim=(0,5.2))
plot.set_title('Relationship between perceived ability of critical questioning & evaluation and moral autonomy', fontsi
ze=18)
plt.legend(loc='lower right')
plt.show()
```

In [ ]:

```

```