Transparency of Artificial Intelligence:

A Discourse Analysis of Dutch Public Opinion Using Q-Methodology

Stefano Sarris

August 2019

A thesis submitted for the fulfillment for the degree of Master of Public Administration

Leiden University, Faculty of Governance and Global Affairs

Under the supervision of **Dr. Brendan Carroll**

Table of contents

Foreword

I would like to thank the participants for taking their time and interest to participate in this study. Without their valuable input, this study would not have been as robust as it is. In addition, my gratitude goes out to those who have supported me throughout the process of this study. The names are too many to mention. First and foremost, I would like to thank my advisor, Dr. Brendan Carroll for his supervision and support. I would also like to express my gratitude to my family and friends for insightful discussions and moral support. I would like to thank Wing Yee, for proofreading this study and providing helpful comments. Finally, I would like to thank Ms. Linah Ababneh for proofreading this study and providing insightful comments.

Abstract

In a time where there is a strong call for AI regulation in the Netherlands and Europe, heated societal debates have proved that it is challenging to determine if, to what extent, and to who AI should be made transparent. The conflicting arguments could pose a barrier for policymakers to devise AI policy that satisfies the ranging interests. On this backdrop, this study set out to ask the question: what are the public discourses of AI transparency in the Netherlands? To answer this question, this study investigated the public discourses on AI transparency on a sample of 31 participants from the Netherlands using Q-methodology. Principal component analysis and varimax rotation registered the presence of three distinct discourses in the sample: i) "no excuses, we demand AI transparency"; ii) "balance the needs for AI transparency"; and iii) "let's reap the benefits of AI." The first discourse was found to be strongly in favor of transparency as means to generate accountability, understanding, and legitimacy; factors which the discourse found necessary to trust AI. This discourse was found to have a strong preference for human-centric AI even when AI is transparent. The second discourse was found to analyze the context of the situation before deciding whether AI should be transparent. This discourse was found to likely make this decision based on a utilitarian approach, considering what would be best for the greater society. The third discourse was found to worry that transparency would affect the performance of AI. Contrary to the "opponents" arguments from the literature, this discourse was found to not be as strongly in opposition to AI transparency. The discourse was found to be open to forms transparency if it does not affect the performance of AI.

Keywords: Artificial Intelligence, Transparency, Public Administration, Public Policy, Q-methodology, Discourse Analysis, Principal Component Analysis

List of figures and tables

Chapter 1: Introduction

The introduction of artificial intelligence (AI) has been marked by many scholars as a milestone in human development. Some scholars are calling it the "AI revolution," a revolution of brain power comparable to the introduction of mechanical power and computing power during the industrial revolution and the digital revolution (Makridakis, 2017). Others are calling it a "disruptive technology'' (Sun & Medaglia, 2019, p. 379), or "a powerful force that is reshaping daily practices, personal and professional interactions, and environments" (Taddeo & Floridi, 2018, p. 751). Regardless of how AI is referred to, one thing is clear: AI is here to stay. Over time AI applications are expected to permeate more and more into the daily lives of people, and in the near distant future AI is expected to be smarter than people (Wang & Siau, 2018).

AI has received fluctuating levels of attention since the 1950s, but its applications have remained mostly confined to academic investigation. This changed over the recent decade due to the affordability of high-performance computing power, the development of improved algorithms, and the introduction of big data (Casares, 2018; Cath, Wachter, Mittelstadt, Taddeo, & Floridi, 2018). These developments facilitated the entry of AI in many domains of everyday life, such as self-driving cars, chatbots, AI judges or doctors, media suggestions on platforms, and facial recognition systems. The examples are voluminous, spanning from the public sector to the private sector, from home development to malicious development. The high number of AI applications which are now present in the daily lives of people have sparked many regulatory questions at the local, national and intergovernmental level.

At the EU level, AI has been discussed since as early as 2016. These discussions have significantly intensified since the Declaration of Cooperation on Artificial Intelligence in April 2018 (European Commission, 2018). Since the declaration, AI has rapidly moved up the policy agenda of the European Union. On 9 May 2019, AI was featured on the Leaders' Agenda of the European Council under the notion of "develop artificial intelligence" (European Council, 2019b). Subsequently on 20 June 2019, AI was prominently headed under one of the four priorities (priority 2: developing our economic base) of the EU's Strategic Agenda for 2019 - 2024 (European Council, 2019a). And on 16 July 2019, Ursula von der Leyen (President-elect of the European Commission) mentioned in her Political Guidelines that she would "put forward legislation… on Artificial Intelligence" in her "first 100 days in office" (von der Leyen, 2019).

Regulatory discussions have heated up in the Netherlands as well. The year 2018 saw a strong call for the development of an AI strategy for the Netherlands. This call culminated in the publication of a report "AI for the Netherlands" by AINED, a Dutch public-private consortium on AI (AINED, 2018; VNO-NCW, 2018). Shortly thereafter on, 21 March 2019, the Dutch Secretary of State, Mona Keijzer, announced that the Netherlands would bring forth a strategic action plan on AI *by the summer* of 2019 (Rijksoverheid, 2019b). Since then, the Netherlands has published a new Digital Strategy '2.0', wherein AI is the highest priority for the year to come (Rijksoverheid, 2019a). The report confirms that the Dutch government will work together with private and public parties to publish a strategic action plan for AI, not during but *after the summer* of 2019 (Rijksoverheid, 2019b).

The discussions at the different levels of governance also focus on the risks and ethical implications of AI. For example at the EU level, the European Commission initiated a High-Level Expert Group on Artificial Intelligence (AI HLEG) which published ethical guidelines for trustworthy AI in April 2019 (AI HLEG, 2019). And as mentioned above, the legislation envisioned to be "put forward" by Ursula von der Leyen in her "first 100 days in office" will be on "a coordinated European approach on the human and ethical implications of Artificial Intelligence" (von der Leyen, 2019).

And in the Netherlands in 2018, the House of Representatives of the Netherlands accepted a motion and held heated debates on the transparency of algorithms and AI that are used by the government (Tweede Kamer, 2018a, 2018b, 2018c). And after the summer of 2019, the Dutch government is aiming to publish a policy plan on AI, public values and human rights (Rijksoverheid, 2019b). In addition, the Dutch government has also started to work together with the private sector and auditors to establish a "Transparency lab for algorithms," which will perform research on how to ensure the explainability and auditability of algorithms (Rijksoverheid, 2019a). And in their report, AINED has called on the Dutch government to establish a social, economic and ethical framework as one of the goals of its national AI strategy (AINED, 2018). Moreover, AI transparency debates are also being held at the local level in the Netherlands. For example, a magazine article by the Association of Netherlands Municipalities emphasized that AI can help municipalities, but that "civil servants should stay in control" and not follow AI blindly (VNG, 2018).

One challenge that is salient in the discussions on risks and ethics of AI is the topic of AI transparency (Cath et al., 2018; AI HLEG, 2019). The transparency problem exists because of the nature of AI: the system is oftentimes labeled as a "black-box" (i.e. a non-transparent) because the complex nature of AI makes it difficult to understand how AI decisions are made (Adadi & Berrada, 2018; Strauß, 2018). In some cases of AI, the system is said to be at such a level of complexity that even experts cannot understand its functioning (Strauß, 2018). The black-box problem is something which is believed to pose major challenges to policymakers (Sun & Medaglia, 2019). The focus is also on transparency because it is often viewed as an important pillar in democratic systems; for example, it is a topic that is frequently linked to *accountability* (Filgueiras, 2016) and *legitimacy* (Eshuis & Edwards, 2013; Schmidt, 2012).

1.1 Research question and hypothesis

This study has been largely motivated by the risks and ethical implications of AI transparency that are considered by policymakers. And as discussed below, policymakers have access to very few studies on the public opinion and/or discourses of AI transparency. This could make it challenging for policymakers to devise policy which is acceptable to the dominant discourses that are present in society. The aim of this study is therefore to unravel the discourses which exist on AI transparency. The case of study will be the Netherlands public opinion of AI transparency. This thesis will investigate the Dutch public opinion of AI transparency through discourse analysis, using Q-methodology. The research question for this thesis is as follows:

What are the public discourses of AI transparency in the Netherlands?

Answers to the research question are first hypothesized based on the theoretical findings in *chapter 2*. The Q-methodology is subsequently used to measure which discourses are empirically present in a sample of 31 participants (see *chapter 3*). The results of the Q-study will first be presented as is in *chapter 4*. To discover alignment with the literature as well as new findings, the empirical results are compared to the hypotheses and analyzed based on the theoretical findings in *chapter 5*. The hypothesized discourses are:

Hypothetical discourse 1: *the proponents discourse.* The so called "proponents" are those who mainly focus on the benefits of AI transparency. They are also expected to worry about the downsides of non-transparency.

Hypothetical discourse 2: *the opponents discourse.* The so-called "opponents" are those who largely worry about the negative impacts that AI transparency might have. They are also expected to reflect on the benefits of non-transparency.

Hypothetical discourse 3: *the context-dependent discourse.* The need for AI transparency is "context-dependent" for this discourse. They are expected to carefully consider the context of a case to determine whether the benefits of transparency outweigh the costs of transparency.

1.2 Knowledge gap

This study can contribute to fill the knowledge gap in two areas. Firstly, only a few studies could be identified which investigate AI discourses. One study was found from Johnson and Verdicchio (2017) which investigates how AI is conceptualized and presented to the general public, with the aim to alter the often-wrongful discourse that the public has on AI. And paradoxically, a paper from Moore and Wiemer-Hastings (n.d.) was found that discusses how AI and computational linguistics applications use discourses to interpret data. After an extensive search, however, the literature on AI discourses was rather barren. There are some studies which investigate public opinion on the use of AI in a variety of domains, but these do not establish discourses. This is not to say that "no such studies exist," but it does illustrate that AI discourses are underdeveloped at present.

Secondly, no AI public opinion study could be found which explored the public opinion of AI transparency. Two studies were found which investigated the Dutch public opinion of AI (Araujo et al., 2019; Verhue & Mol, 2018). Yet these studies were not found to investigate the Dutch public opinion of AI transparency. In the EU, Special Eurobarometer 460 (TNS, 2017) on "attitudes towards the impact of digitization and automation on daily life" extensively investigated AI, but not in the context of transparency. And in the U.S., a study by Smith (2018) was done on the "Public Attitudes Toward Computer Algorithms," but this study also did not investigate AI transparency. Another study in the U.S. by Zhang and Dafoe (2019) on "Artificial Intelligence: American Attitudes and Trends" did pose some transparency related questions to their participants, but did not explicitly treat AI transparency in their results.

1.3 Academic and societal relevance

This research contributes to the academic literature by providing new insights into discourses and public opinion on AI transparency. As identified above, this direction is an under-explored area in the literature. Given the relevance of AI transparency to policymakers, the findings of this study can contribute directly to the fields of AI Governance, Public Administration, AI Ethics, and related fields. The findings of this study can also contribute to the more technical scholarly fields of AI, one example would be the field of XAI (explainable AI), the broader field of machine learning, and the more specialized field of deep learning.

This research can also contribute to the methodological development to study AI discourses. Although one study proposes the integration of a specific form of AI, neural networks, to improve the Q methodology (Eghbalighazijahani, Hine & Kashyap, 2013), this research is the first in the field that uses the Q methodology (the research method of this paper) to investigate public discourses on AI. In that sense, it can serve as a template for future Q-methodology studies on AI.

Regarding societal relevance, the outcome of this study could feed into the real-world policy developments which are currently taking place in the Netherlands and in the EU. It can give politicians, policymakers and other civil servants an idea regarding which public discourses exist on AI transparency. The results of this study could facilitate stakeholder consultations as AI policymakers are contemplating, formulating, implementing, or evaluating AI policies and legislation. Moreover, the results could assist non-public sector stakeholders (such as in the medical sphere or at educational institutions) in creating AI systems which are acceptable to the varying discourses.

1.4 Methods and data collection

The Q-methodology was deployed to investigate the Dutch public opinion on Al transparency in the form of discourses. For this method, a keyword search was used to gather a variety of perspectives in statement-form from Dutch written sources. This resulted in a concourse of 233 statements. The concourse was classified and narrowed down to 54 statements (Q-set) using a self-developed analytical framework on AI transparency (modelled after a conceptual framework developed in *chapter 2*) and pre-defined selection criteria from the literature. The Q-set was sorted by 31 participants on a quasi-normal distribution grid based on a scale from -5 (least how I think) to +5 (most how I think). In addition, participants were able to comment on their ranking choices.

1.5 Data analysis

Factor analysis was performed on the 31 statement sorts (also called Q-sorts) using the PQMethod software (version 2.35) by Peter Schmolck (2014). Principal component analysis (PCA) and automated varimax rotation revealed that 3 factors produced the most robust quantitative results. The 3 factors were first described as discourses on the basis of the empirical findings. The discourses were subsequently analyzed and discussed to identify met and unmet expectations from the hypotheses, and to discover the potential meaning of AI transparency for the specific discourse.

1.6 Thesis outline and structure

This thesis is composed of six chapters. **Chapter 1** introduces the reader to the study rationale by discussing the research problem and question, the relevance of this study, the research methodology and mode of analysis. **Chapter 2** (literature and theory) introduces AI and transparency as separate concepts, and then dives deeper into the concept of AI transparency. At the end of the chapter, discourses on AI transparency are hypothesized on the basis of expectations from the literature. **Chapter 3** (methodology) will explain how the Q-methodology is deployed to investigate the research question of this study. Moreover, it will explain how the conceptual framework (from *chapter 2*) is operationalized and how the results are analyzed. **Chapter 4** (empirical findings) strictly reports on the empirical findings of this study. **Chapter 5** (analysis and discussion) compares the empirical results with the theoretical expectations of *chapter 2* and it analyzes the results with the theory of *chapter 2* to discover the potential meaning of AI transparency for the specific discourse. **Chapter 6** (conclusion) will summarize the thesis and reflect and answer the research question in a concise manner. Moreover, it will briefly discuss some key take-away messages, such as the strengths and limits of this study, and future research suggestions. The outline and steps taken to answer the research question of this thesis can be found in the research framework below (*figure 1*).

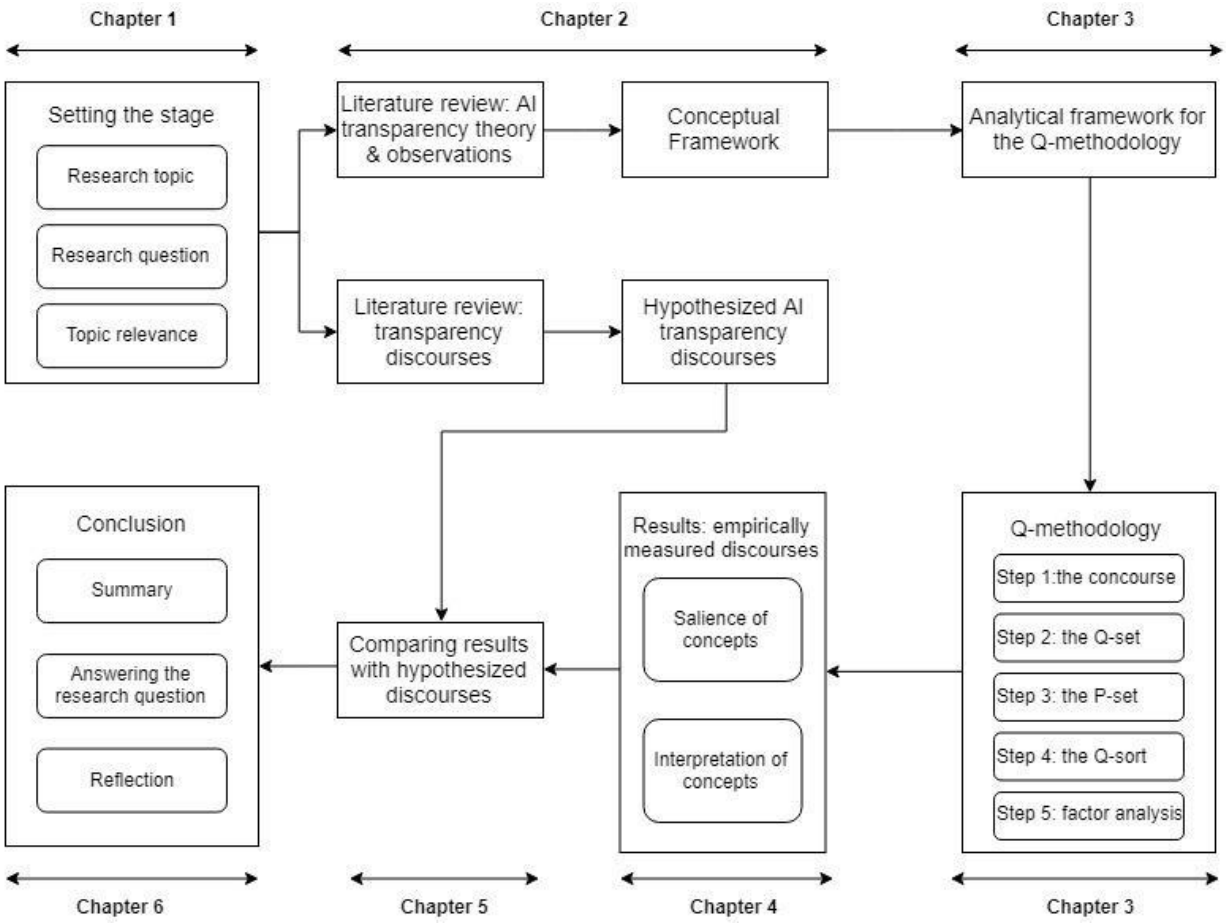


Figure 1: Research framework illustrating the steps taken to answer the research question.

Chapter 2: Theory

The purpose of this chapter is to set up the theoretical underpinnings that can support a Q-methodology study on the Dutch public discourses of AI transparency. To make this possible, the first goal of this chapter is to review the theory on AI transparency to arrive at a conceptual framework. This conceptual framework will support the classification and selection of statements for the Q-methodology (see *chapter 3*). The conceptual framework will serve as an analysis tool to investigate the empirical discourse outcomes in *chapter* 5. The second goal of this chapter is to hypothesize, based on the extant literature, which discourses on AI transparency are likely to exist. These hypothetical discourses will be compared with the discourses that will result from the Q-methodology study (see *chapter 5*).

The outline of this chapter is as follows. The first section of this chapter will briefly introduce artificial intelligence (AI) to set the context of this study. The second section will turn to the conception of AI, to arrive at a workable definition. The third section will focus on the conception of transparency and subsequently AI transparency. The fourth section will discuss the dominant factors that, according to the literature, can contribute to making AI transparent. The fifth section turns to the concepts surrounding the levels of transparency, including full transparency, limited transparency, black-boxes, and opacity. The sixth section identifies the possible effects of transparent and/or non-transparent AI. In the seventh section, a conceptual framework is established on the basis of the factors of AI transparency, levels of AI transparency and effects of AI transparency. Finally, in the eighth section the discourse literature and the AI and computer mediated transparency literature is reviewed to develop hypothetical discourses on AI transparency.

2.1 A brief timeline perspective of Artificial Intelligence

In 1950, Alan Turning asked the question: “can machines think?” and contemplated whether machines could one day learn (Turing, 1950, p. 433). In the concluding section of one of his papers, Turing (1950) wrote:

> we may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. (p. 460)

More work shortly followed. In 1956, it was John McCarthy, by some called the father of AI, who first mentioned the term AI in a research proposal for a summer programme at Dartmouth College (Rajaraman, 2014). McCarthy had previously crossed paths with Turing in 1948, and he worked with Claude Shannon who (among others) eventually wrote papers on machine intelligence and teaching computers chess (Rajaraman, 2014).

Along the timeline of AI, and even well before 1950 (Strauß, 2018), there were many AI landmark developments; to name them all would be beyond the scope of this study. However, one benchmark event happened in 1997 - it was the year when IBM’s *Deep Blue* defeated the world chess champion Gary

Kasparov in a chess match (Jarrahi, 2018). It marked the beginning of an era where AI, as once contemplated by Turning and Shannon, could artificially perform intelligent functions at a high level, such as playing chess.

Today, in the year 2019, AI has been further extracted from the world of abstractions. This development has been widely attributed to the development of more sophisticated algorithms, affordable powerful computing power; the availability of big data, and the growing interconnectivity of systems and utilities (e.g. see Casares, 2018; Cath, Wachter et al., 2018; Government Office for Science, 2016). AI applications are permeating into everyday life; they are brought by the private sector into households (e.g. self-driving cars and personalised advertisements), in the health sector (Pesapane, Volonté, Codari, & Sardanelli, 2018), at educational institutions (Kaplan & Haenlein, 2019), in government (Helbing, 2018), and a number of other domains.

The trend of AI growth is expected to continue in the future. Scholars are now making the case for even more sophisticated applications, which would allow AI to correct itself and provide future solutions (Kirkpatrick et al., 2017). Google director Ray Kurzweil even goes as far as stating that machines will outperform human minds by the year 2030 (Helbing, 2018). And at the same time, scholars are contemplating whether AI means the end of democracy as we know it (Helbing, 2018). But what is AI precisely? The following section will further dive into the concept of AI.

2.2 Defining Artificial Intelligence

There have been many attempts to define AI, for an overview of a list of AI definitions, one could consult Buiten (2019). As Gasser and Almeida (2017) explain, there is "no universally accepted definition of AI... the term AI is often used as an umbrella term to refer to a certain degree of autonomy exhibited" (Gasser & Almeida, 2017). One reason why a definition for AI is missing is because it "is not a single technology, but rather a set of techniques and sub-disciplines" (Gasser & Almeida, 2017). This message resonated throughout the literature review. This makes it a challenging task to arrive at a workable definition for AI that satisfies all AI types. Therefore, the approach was taken to identify which attributes are essential for the functioning of AI (for conception and the use of attributes, see Toshkov, 2016, p. 89). The main AI attributes discussed below are: algorithms, big data, computing power, cognition, automation, and intelligent thought.

2.2.1 Algorithms

Alike AI, algorithms do not have a widely acknowledged definition (Brkan, 2019). Cormen (2013, p. 1) defines algorithms as "a set of steps to accomplish a task." And he defines computer algorithms more specifically as "a set of steps to accomplish a task that is described precisely enough that a computer can run it." One of the core elements which permit the functioning of AI are its underlying algorithms. Because AI requires computing power, this paper uses the above definition of *computer algorithms* as a starting point for the definition of AI algorithms. In addition, AI models and AI algorithms are often referred to as being 'complex' (e.g. see Strauß, 2018). Therefore, in this paper, AI algorithms are sometimes referred to as *complex algorithms*.

2.2.2 Big data

There is also no "universally agreed definition" for big data (Yau & Lau, 2018, p. 1). Some agreement exists on the four factors which make up big data, which are volume, velocity, variety, and complexity (Desouza & Jacob, 2017). Others express that big data is "digital.... [which] means that huge amounts of data are available to anyone in the world over the internet" (Johnson, Denning, Delic & Sousa-rodrigues, 2018, p. 2). Or that big data is a "new landscape of the data ecosystem… a wide spectrum of datasets with varying characteristics" (Yau & Lau, 2018, p. 2709). Big data and AI have plentiful of overlaps, which is in part because they both use datasets, algorithms, and some form of computing power to function (Strauß, 2018). But they differ in the sense that AI can perform intelligent, automated, and cognitive functions (Strauß, 2018). For a full overlap of the similarities and differences between AI and big data, one could consult Strauß (2018). For the purpose of this paper, I will keep the definition of big data rather simple, in that they are *large datasets* (Pesapane et al., 2018).

2.2.3 Automation, Cognition and Intelligence

As mentioned above, AI has automated, cognitive, and intelligent capacities. Automated decision-making can be broadly defined "as taking a decision without human intervention" (Brkan, 2019, p. 3). With cognitive capacities, this study refers to functions such as thinking and learning (Strauß, 2018) and recognition.

Intelligence is also a broad term. In Merriam-Webster (2019) it is referred to as:

> the ability to learn or understand or to deal with new or trying situations... the skilled use of reason [and]... the ability to apply knowledge to manipulate one's environment or to think abstractly as measured by objective criteria (such as tests) [and]... the act of understanding [and]… the ability to perform computer functions.

Intelligence is also a measure in terms of how smart someone or something is (e.g. think about IQ scores). Strong and weak AI classifications also exist. When AI is designed to execute specific or single tasks it is classified as 'weak' or 'narrow' AI (Wang & Siau, 2018). A classification of 'strong' or 'Artificial General Intelligence (AGI)' is used when AI can use intelligence to multitask, address problems, or when it is self-aware, and when it can expresses genuine intelligence (Gasser & Almeida, 2017; Wang & Siau, 2018). All AI applications that are presently in use are classified as weak AI. Strong AI is anticipated to be decades or centuries away (Wang & Siau, 2018).

Automated, intelligent and cognitive capacities can be recognized across the AI domain of *machine learning* (ML), currently the most dominant field for AI developments (Calo, 2017). Machine learning is a broad term for a class of computational techniques where algorithms can learn from data (Anastasopoulos & Whitford, 2018). Machine learning can be categorised into *unsupervised learning* and *supervised learning*. Unsupervised learning is used to uncover patterns within unclassified data, and supervised learning is used to create a model based on provided data which can predict the outcome of new data (Anastasopoulos & Whitford, 2018). An advanced sub-category of machine learning is *deep learning*, which is a set of algorithms that can automatically make predictions, extract features and identify patterns from large (unsupervised) data sets without human involvement (Jan et al., 2019; Pesapane et al., 2018).

Deep learning is capable of performing such functions because its fundamental structure is based on a *deep neural network* (see *figure 2*); these networks are inspired by the cognitive functioning of the human brain (Hegelich, 2017). A deep learning model is comprised of several layers (input, hidden, and output) of data-processing points (like neurons in the brain) which are webbed together in a non-linear network (Hegelich, 2017). It operates by transforming a given input through a process of re-iteration, where connections between neurons are re-weighted until a desired output (or as close as possible by the model) is reached (Hegelich, 2017).

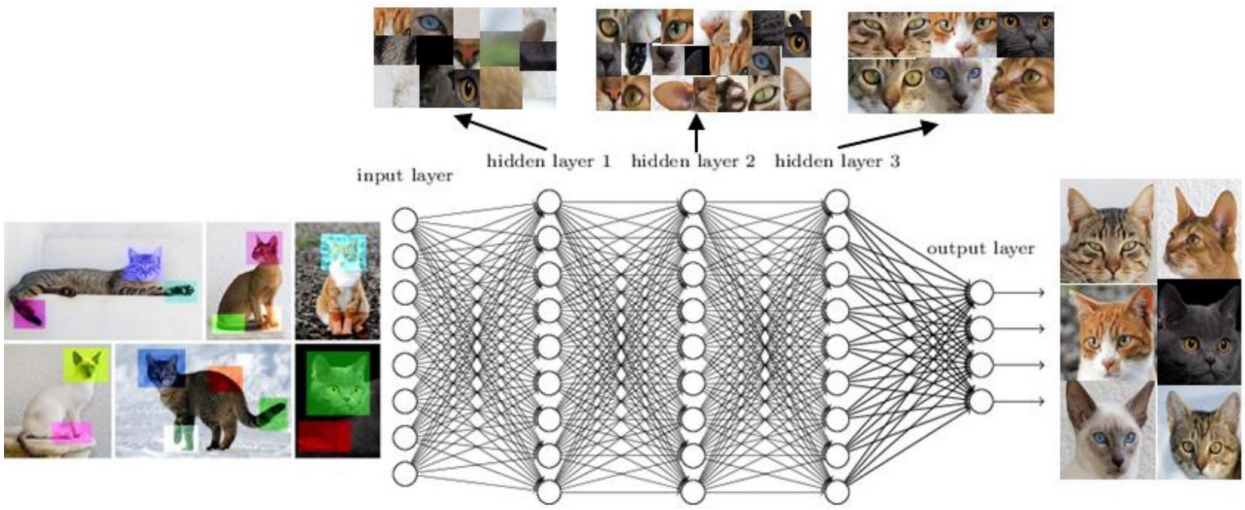


Figure 2. Example of image recognition based on neural network layers (adapted from Strauß, 2018)

2.2.4 Computing power

The combined developments in large datasets, algorithms and high computing power have been primary contributing factors to enable the everyday use of AI (Casares, 2018). Computing power is essential because AI often requires high computing power to function (Casares, 2018; Strauß, 2018). And computing power has become increasingly more affordable - the number of computations per unit of energy has doubled roughly every 1.57 years over the past seventy years (Casares, 2018). High computing power can thus be understood as one of the main factors which AI needs to function, and its affordability has permitted AI to become more mainstream. Computing power in this paper will be defined as *the number of computations per unit of energy*. High computing power in this case would mean a (relatively) high number of computations per unit of energy.

2.2.5 Towards a workable definition

As discussed, algorithms, large datasets, computing power, automation, intelligence and cognition are all factors play a role in the functioning of AI. These factors also come to the fore in an AI description by Pesapane et al. (2018). The description by Pesapane et al., (2018) is a workable starting point, but it misses the notion of automation:

> AI is a branch of computer science dedicated to the creation of systems that perform tasks that usually require human intelligence with different technical approaches. The term AI is used to describe computer systems that mimic cognitive functions, such as learning and problem-solving.

> These systems are currently based on artificial neural networks, which are flexible mathematical models using multiple algorithms to identify complex non-linear relationships within large datasets, nowadays known as big data. (pp. 745-746)

Based on the above findings of AI attributes, this paper will define AI as: *a set of complex algorithms that require high computing power and access to large datasets to perform automated, cognitive or intelligent duties.*

2.3 Defining AI Transparency

The aim of this section is to arrive at a workable concept for AI transparency. The focus will first be on transparency, the parent concept of AI transparency. Considering the definition for transparency, and the attributes of AI transparency, a concept for AI transparency will be established.

2.3.1. Transparency

To start off, the concepts of openness and transparency are frequently used in an interchangeable manner in the literature, but they are sometimes treated as stand-alone concepts (e.g. see Meijer, Hillebrandt, Curtin, & Brandsma, 2010; Moore, 2018). Meijer et al. (2010) refer to openness as "open access to decision-making arenas," and transparency as "open access to government information." In this paper, the concept of openness will be treated as part of the concept of transparency, and not as a distinct stand-alone concept

There is no widely accepted definition for the concept of transparency. This is possibly because transparency is discussed in a broad range of academic disciplines. This includes the field of political science, governance, and public administration, as well as that of AI and related disciplines (such as on algorithms). To illustrate the academic richness of the concept, a study performed by Cucciniello, Porumbescu, & Grimmelikhuijsen (2017) found that 177 peer-reviewed works and 10 monographs were produced solely on the topic of *government transparency* between 1990 and 2015.

In Merriam-Webster (2019b) transparency is defined as "the quality or state of being transparent." Transparent is defined in Merriam-Webster (2019c) as "free from pretense or deceit… [and] easily detected or seen through… [and] readily understood… [and] characterized by visibility or accessibility of information especially concerning business practices."

The many definitions in Merriam-Webster are in line with Kosack and Fung's (2014, p. 67) remark that transparency has "multiple meanings, as well as multiple rationales, purposes, and applications." Turilli and Floridi (2009, p. 105) define transparency as "information visibility… in particular… to the possibility of accessing information, intentions or behaviours that have been intentionally revealed through a process of disclosure." Cucciniello et al. (2017) identified two larger definition categories for transparency, one emphasises the "flow of information," and another "information availability."

Buiten (2019) explains that the concrete interpretation of transparency depends on the context and purpose for which it is used. Perhaps a definition of transparency ultimately depends on the transparency of *what*, to *who* and *when*. What as in, what is being made transparent? When as in the inputs, the

process, or the outcomes (Buiten, 2019)? And who as in, who is the receiver and provider of transparency?

In this paper the concept of transparency will be kept broad as: *the availability, visibility, and accessibility of information flow.* This broad definition leaves the what, when, and who category rather open. Availability, visibility and accessibility are included as broad factors which can contribute to transparent information.

Now that we have a working definition for transparency, the following section will turn to defining AI transparency, the subject of this paper.

2.3.2 AI transparency

To conceptualize AI transparency, it is first necessary to discover its attributes (Toshkov, 2016, p. 89). To identify the attributes (called 'factors' from here onwards) of transparency, I started with the basic question: what makes AI transparent? In the section below, five salient factors are identified that, according to the literature, contribute to making AI transparent. These are: explainability, interpretability, traceability, auditability, and communication. These factors probivde a workable basis to conceptualize AI transparency. By integrating the parent concept of transparency, this paper defines AI transparency as *enhanced explainability, interpretability, traceability, auditability or communication which makes AI information flow more available, visible or accessible.*

2.3.3 From a definition towards a conceptual framework on AI transparency

The following sections will specifically examine the existing literature and theories surrounding AI transparency. The next section will first examine the factors which enable AI to be transparent. The subsequent section will identify the levels of AI transparency. And the section thereafter will identify the effects of AI transparency. The combined findings of these three sections will culminate into the development of a conceptual framework on AI transparency.

2.4 Five factors of AI transparency

The literature on AI and related fields were analyzed to map the dominant AI transparency factors. Five dominant factors of AI transparency were identified in the literature. These are: *explainability, interpretability, traceability, auditability, and communication.* These factors are not exhaustive, but they were observed to be the most dominant factors in the literature. For example, one factor which is not discussed is *explicability.* This is because it appeared to be a parent concept of *explainability, traceability, auditability,* and *communication* (AI HLEG, 2019). Another concept which is not included is that of *information,* which plays a role in all five dominant factors. The concepts of *understandability* (e.g. see de Laat, 2018 or Lepri, Oliver, Letouzé, Pentland, & Vinck, 2018) and *comprehensibility* (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016) were omitted because they overlap largely with interpretability and explainability. More factors could be included in as factors of *AI transparency* in future works.

2.4.1 Explainability

There is an entire scholarly field called XAI (eXplainable AI) which is evaluating whether we can make AI explainable (AI HLEG, 2019; Miller, 2019). In some works, explainable AI is defined as "artificial

intelligence and machine learning techniques that can provide human understandable justification for their behaviour" (Ehsan, Tambwekar, Chan, Harrison, & Riedl, 2019, p. 1). The definition mentions 'techniques' because there are several ways in which AI could be made explainable. Three larger classes of AI explainability exist according to Ehsan et al. (2019): the first is that data and the workings of the system are presented *as is*; the second would be to add a form of rationale in natural language, which are fitting to the context; the third form is full communication in human form. In some works, explainability is at the heart of the definition for AI transparency. For example, in Ras, van Gerven, & Haselager (2018, p. 5) "transparency refers to the extent to which an *explanation* makes a specific outcome understandable to a particular (group of) users." A similar interpretation of transparency can be found in Lepri et al. (2018). The factor of interpretability is closely related to the explainability of AI. The difference between the two for this thesis will be discussed in the section below.

2.4.2 Interpretability

A good distinction between the concepts of interpretability and explainability can be found in Mittelstadt, Russell, & Wachter, (2019) who explain that 'interpretability' is defined by the ability of humans to understand a decision; whereas 'explanation' refers to the exchange of information about a process to stakeholders (Mittelstadt et al., 2019). According to Lepri et al. (2018), two types of interpretability exist: "the first one relates to *transparency*, that is *how does the model work,* [and] the second one consists of *post-hoc interpretations*, that is *what else can the model tell*." Interpretability and transparency are also linked through the 'interpretability problem' which posits that certain AI applications are *intrinsically opaque* by design, which makes them challenging to interpret (Lepri et al., 2018). The difference between interpretability and explainability that can be discerned here is that interpretability is more about understanding AI decisions, and explainability about information exchange. Some conceptions on explainability in the section above do include *understanding*, but in this paper, *understanding* will fall under interpretability. This paper will define interpretability and explainability as found in Mittelstadt et al. (2019), where interpretability is *the ability of humans to understand an AI decision*; and explainability refers to *the exchange of information about an AI process to stakeholders.*

2.4.3 Traceability

In AI HLEG (2019) the factor of traceability means that datasets, decision-processes (including gathering and labelling of data and the use of algorithms) is documented with the goal to improve transparency. The linkage between transparency and traceability can also be found in Buiten (2019, p. 14), who writes that "[t]ransparency means tracing back how certain factors were used to reach an outcome in a specific situation." In this paper, traceability will refer to *the ability to trace back, step by step, how a certain AI outcome came to be*. This can relate to the outcome, process, or inputs of AI.

2.4.4 Auditability

Traceability is said to be a 'facilitator' of auditability (AI HLEG, 2019). However, traceability is not a necessary condition of auditability. For example, audits could also "reverse engineer" AI processes when the systems inputs and outputs are visible (de Laat, 2018). A variety of examples of auditability as means to make algorithmic processes more transparent can be found in Lepri et al. (2018), Sandvig, Hamilton, Karahalios, & Langbort (2014), and Zarsky (2016). In AI HLEG (2019) auditability is more broadly referred to as the ability to audit "algorithms, data and design processes" of an AI system. In this paper, auditability will refer broadly to *the ability to have oversight, inspect, or audit AI.*

2.4.5 Communication

The factor of communication was introduced by the AI HLEG (2019, p. 18). With communication it is meant that: 1) "AI systems should not represent themselves as humans to users; humans have the right to be informed that they are interacting with an AI system," and that 2) the "AI system's capabilities and limitations" are communicated. In this paper, communication of AI is defined as *the ability for persons to identify and set expectations regarding an AI system when they interact with one another.* An example could be a chatbot on a website. To fulfill the communication requirement, the person would have to be notified if the interaction was in fact with a chatbot. And the second criteria would be met if the chatbot mentions what it can and/or cannot do.

2.5 Levels of AI transparency

The previous section discusses the factors which enable AI to be transparent. This section will specifically discuss the main levels of AI transparency that are frequently mentioned in the literature. The concepts that are addressed here are: black-boxes, opacity, limited transparency and full transparency.

2.5.1 Non-transparency: the black-box problem and opacity

The literature on AI transparency has two specific concepts for non-transparency: *black-boxes* and *opacity*. The black-box problem is one of the focal points in the literature on AI transparency (e.g. see Adadi & Berrada, 2018, Casalicchio, Molnar, & Bischl, 2019, Samek, Wiegand, & Müller, 2017). The problem is sometimes referred to as the need to "open the black-box" (Lorscheid, Heine, & Meyer, 2012; Samek et al., 2017). An AI system is deemed to be "an opaque black-box [when it provides] users scarce visibility about the underlying data, processes, and logic that leads to the system's decisions" (Rossi, 2019, p. 129). Another paper writes that black-boxes are "nested nonlinear structures [which] make them highly non-transparent, i.e., it is not clear what information in the input data makes them actually arrive at their decisions" (Samek et al., 2017, p. 1). The problem is sometimes so critical that even AI experts can no longer decipher AI processes (Strauß, 2018).

The black-box problem also comes to the fore in other works on transparency. For example, Schmidt (2012, p. 3) refers to the EU policy-making process as "the black-box of EU governance." (Hale, 2008, p. 76) refers to transparency and the "black-box of politics." The concept of a black-box originates from systems theory and it "resembles a system viewed in terms of inputs causally related to outputs, without knowledge of the system's internal workings" (Gössling, Cohen, & Hares, 2016, p. 86). In this paper, the concept of a black-box will be broadly defined as *a process of a system which is opaque*. For example, in policy a black-box would be the opacity of a policy process. And in AI it would refer to the opacity of an AI decision-making process.

But what is opacity? The terms opacity and black-boxes in the AI literature are often used interchangeably. However, there are slight differences. Opacity more generally refers to something that is non-transparent. Whereas a black-box often specifically refers to the process of an AI system. In this paper, the meaning of *opacity* will be adapted from Lepri et al. (2018, p.620), who calls it a *"lack of transparency. "*

2.5.2 Limited transparency and full transparency

Different gradations exist on the level of AI transparency. The three main gradations are: opacity/black-boxes, limited transparency and full transparency. Moving away from the discussion on opacity and black-boxes, the differences between limited transparency and full transparency are widely discussed in de Laat (2018). In this paper, full transparency and limited transparency at the principle level will dependent on the presence or absence of the five factors of transparency that are discussed above. As discussed, these factors can enable an AI system to be transparent. In this paper, limited and full transparency can further be viewed from three angles: transparency of the system (what is made transparent), transparency to certain actors (to who it is made transparent), the temporal element of transparency (when it is made transparent). For example, limited transparency could be that an AI system used by intelligence agencies is made explainable to its staff, but is not explainable to the outside world (to who it is made transparent). Another limited transparency example is that a data-set of a system is auditable, but that the algorithm is not auditable (what is made transparent). A final limited transparency example is that data is only made available during an audit, and that the data is kept opaque before and after the audit (when it is made transparent).

2.6 Effects of AI transparency

This section discusses the theories and empirical observations related to the effects of AI transparency. The effects discussed below are not exhaustive, but they are those which according to the literature are most salient to AI transparency. The most salient effects of AI transparency that were found are: accountability, legitimacy, trust, fear, fairness, privacy, trade-offs, and perverse effects.

2.6.1 Accountability

The effect of AI transparency on accountability is frequently discussed in the literature. A positive relationship between transparency and accountability can be found for example in Ras et al. (2018, p. 5), who writes that "transparency is normally a precondition for accountability." This relationship is also echoed in de Laat (2018) and Lepri et al. (2018).

It should be noted here that transparency is not a necessary requirement of accountability. Accountability can also be attained when certain elements of AI systems are not transparent (Lepri et al., 2018). The same goes for transparency in other domains. For example, in Cucciniello et al. (2017) transparency was found to have positive, mixed and no effect results on government accountability. And Papadopoulos (2010, p. 1034) writes specifically that "transparency and access to information… are no substitute for genuine accountability mechanisms… even though transparency and publicity are often cited as a remedy for accountability problems, although necessary, they are not sufficient."

Ras et al. (2018, p. 5) defines accountability as "the extent to which the responsibility for the actionable outcome can be attributed to legally (or morally) relevant agents (governments, companies, experts or lay users, etc)." In this paper, accountability will be more broadly defined as *knowing "who is responsible"* (Risse & Kleine, 2007, p. 73).

2.6.2 Legitimacy

Veale and Brass (2019) conclude in their paper that algorithmic decision-making can give rise to concerns regarding legitimacy. The authors write that transparency can have an effect on public legitimacy, but that the outcome depends per type of policy, especially when trade-offs exist (Veale & Brass, 2019). The mixed outcome of transparency on legitimacy also comes to the fore in Cucciniello et al. (2017), the study reports that two thirds papers found on transparency and legitimacy had mixed results (positive and negative effects) and that one third had positive results only.

Some scholarly works analyze legitimacy from a systems theory approach. In these papers, transparency is often considered to be an enabling factor of "throughput legitimacy" (Eshuis & Edwards, 2013; Risse & Kleine, 2007; Schmidt, 2012). Throughput legitimacy specifically "concerns the quality of the decision-making process itself," it focuses on the process that takes place between the input and output phase of a given system (Risse & Kleine, 2007). Risse and Kleine (2007) see an indirect relationship between transparency and legitimacy, they write that it is transparency which ensures that actors can be held accountable, and that accountability in turn generates legitimacy. In Schmidt (2012), the author writes that transparency facilitates throughput legitimacy because it permits the public to have access to information. Beyond legitimacy of the process, legitimacy could also refer to the use of data or the output of a system. For example, one study reports that transparency could negatively affect legitimacy because there is a chance of producing (output) a false negative (e.g. see Veale & Brass, 2019).

Eshuis and Edwards (2013) highlight that many definitions for the concept of legitimacy exist. For the purpose of this study I will use a broad interpretation from Eshuis and Edwards (2013, p. 1070) that legitimacy refers to *the "justifiability of a power relationship."* A power relationship is referred to because the decisions that an AI system makes are a form of power exerted on humans. And legitimacy can be earned when this power relationship can be justified.

2.6.3 Trust

Trust is said to be one of the major obstacles which is holding back the development of AI (Rossi, 2019; Siau & Wang, 2018). Many conceptions of trust exist, but there is no universal agreement. This is partially the case because trust can be context-dependent (Mabillard & Pasquier, 2016). Siau and Wang (2018) interpret trust from three angles:

> (1) a set of specific beliefs dealing with benevolence, competence, integrity, and predictability (trusting beliefs); (2) the willingness of one party to depend on another in a risky situation (trusting intention); or (3) the combination of these elements. (p. 47)

However, the authors note that the concept of trust in the interaction between humans and machines differs (Siau & Wang, 2018).

But then what is a workable concept of trust? According to Grimmelikhuijsen, Porumbescu, Hong, and Im (2013, p. 9) an often cited definition across disciplines comes from Rousseau, Sitkin, Burt, & Camerer (1998) that "trust is a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another." Inspired by Rousseau et al. (1998), in this paper AI trust will be defined as: *the willingness or ability to accept the intentions and behavior of AI.*

The effect of the five factors of transparency to trust is echoed across the AI literature (AI HLEG, 2019; Buiten, 2019; Cath et al., 2018; Helbing, 2018; Lepri et al., 2018; Ras et al., 2018; Riedl, 2019; Samek et al., 2017; Siau & Wang, 2018). This is also the case for political science and related disciplines (Brown, Vandekerckhove, & Dreyfus, 2014; Cucciniello et al., 2017; Grimmelikhuijsen et al., 2013; Mabillard & Pasquier, 2016). A study on secondary data undertaken by Mabillard and Pasquier (2016, p. 84) could not find a positive relationship on whether "greater transparency lead[s] to greater trust" in the government. In Cucciniello et al. (2017) findings also varied, a positive relationship was reported in seven studies, a negative relationship in four, a mixed relationship in six, and no effect in one study.

2.6.4 Fear

Another salient effect is that between the level of AI transparency and fear. Winfield and Jirotka (2018, p. 5) mentions that "it is well understood that there are public fears around robotics and artificial intelligence… some are grounded in genuine worries over how the technology might impact, for instance, jobs or privacy." The challenge of fear for AI opacity can be best summarized as follows: "the opacity of [AI] reinforces concerns about the uncontrollability of new technologies: we fear what we do not know" (Buiten, 2019, p. 3).

An example of the effect of transparency on fear can be found in a paper on government communication and transparency from Fairbanks, Plowman, and Rawlins (2007). Fairbanks et al. (2007) found in interviews with government officials that being transparent could:

> end the fear that decisions on government agencies have been made as a result of undue political or industry influence because the process is open to the public… [which promotes a] better, smoother, more friction free society where you don't have everybody sitting around gnashing their teeth, thinking the worst of institutions… [and that it] creates a feeling of trust in your government. (p. 28)

The last notion somewhat entangles the concept of fear with the concept of trust. That is because the two concepts have associations with one another, for example, de Cremer (1999, p. 53) mentions that trust has "an effect on people's experiences of fear."

Transparency has not only been found to lessen fear, it can also be a factor that generates fear. For example, (Stiglitz, 1999, p. 9) mentions that an incentive for secrecy could for example be related to "fearing that openness allows demagogues to enter the fray and to sway innocent voters." And Fairbanks et al. (2007) found that some government officials would not be transparent because of the fear of providing misinformation which could be a "career ender."

The question therefore remains whether transparency would be able to stymie some of the fears in AI, such as: the fear for manipulation (Helbing, 2018); the fear of undermining democracy and freedom of speech (Helbing, 2018); the fear of superintelligence (Burton et al., 2017; Makridakis, 2017); the fear of harm (Samek et al., 2017); or the fear of being discriminated against (Buiten, 2019).

2.6.5 Fairness

The effect of AI transparency on fairness is also salient in the literature. The challenges which often come to the fore are *biases* which cause *discrimination* and *prejudice* (Brkan, 2019; Calo, 2017; de Laat, 2018; Gasser & Almeida, 2017; Lepri et al., 2018; Riedl, 2019; Rossi, 2019; Samek et al., 2017; Strauß, 2018). Other challenges exist as well. For example, Zarsky (2016) investigates whether transparency could alleviate the challenges of: "(a) unfair transfers of wealth; (b) unfair differential treatment of similar individuals; and (c) unfair harms to individual autonomy." He argues that in some cases transparency would exacerbate unfairness (e.g. when a special interest group uses the algorithmic information to impact decisions, when trade secrets are revealed), and that it can improve fairness (e.g. when it can alleviate bias and discrimination).

The aim here is not to dive into all the details of *AI fairness*, which is a scholarly sub-field of AI potentially as large as that of *AI transparency*; the aim is to arrive at a workable concept for AI fairness. Brkan (2019) interprets fairness as having two dimensions: procedural fairness, and substantive fairness; where procedural fairness focuses on that decision procedures do not deviate in comparable or similar situations; and substantive fairness focusing on prevention of discrimination. The AI HLEG (2019) refers to substantive fairness as focusing on

> equal and just distribution of both benefits and costs, [as well as]... bias, discrimination and stigmatisation;" and procedural fairness on "the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them. (p. 12)

More general conceptions (e.g. Lepri et al., 2018, p. 615) refer to fairness "as the lack of discrimination or bias". The AI HLEG (2019) refers to proportionality, in that it requires the balancing between the rights and interests of deployers (e.g. confidentiality, and intellectual property) and the rights and interests of the user. In addition to confidentiality and intellectual property, Zarsky (2016) also mentions that transparency might affect competitiveness and incentives for innovation.

Fairness of AI in this paper will remain broad as *free of bias, discrimination and prejudice, and balancing the rights and interests of users and deployers.* The effect of transparency on fairness, as identified above, can be positive and negative. For example, it could be negative when transparency would lead to reveal trade secrets, and it could be positive when transparency can lead to the prevention of discrimination.

2.6.6 Privacy

Privacy broadly speaking "is our right to live our lives without any external involvement" (Janssen & van den Hoven, 2015, p. 363). As de Laat (2018) mentions, the privacy argument is often used a "counter-argument" to AI transparency in the AI literature. The basic premise of the argument is that AI transparency could lead to the leakage of data into the public sphere. The leaked datasets would then be used for purposes other than what was intended and affect personal data privacy (de Laat, 2018; Mittelstadt et al., 2016; Ras et al., 2018). This is particularly problematic because transparency could then be subject to breaching various rights in the General Data Protection Regulation (GDPR, regulation EU 2016/679) in the EU (EU Parliament & Council of the EU, 2016; European Commission, n.d.). On the other hand, privacy could also relate to making the use of personal data transparent to the data subject only. This is described in article 15 of the GDPR as "right of access by the data subject" (EU Parliament & Council of the EU, 2016).

This paper will refer to the effect of transparency on privacy in terms of leakage of personal data, and as well the right to access to personal data.

2.6.7 Trade-off effects

Another effect which comes to the fore in the literature is that AI transparency can have trade-off effects. One example would be that greater AI transparency requirements can translate to greater costs (e.g. see Buiten, 2019; Zarsky, 2016). A prominent example in the literature is the trade-off between AI transparency and AI performance (such as the level of accuracy, automation, and capacity of a system).

De Laat (2018) writes that there is a "tension" in the relationship between accuracy and interpretability. The challenge is that AI models are complex and "inherently opaque" which enables accuracy but that this "pushes interpretability into the background" (de Laat, 2018). Lepri et al. (2018, p. 620) writes that the "interpretability problem" can be averted by "using alternative machine learning models that are easy to interpret by humans, despite the fact that they might yield lower accuracy than black-box non-interpretable models." Zarsky (2016, p. 129) also mentions that "various forms of disclosure [is] possibly at the price of simplifying the automated process and compromising its accuracy."

Zarsky (2016, p. 121) moreover refers to a trade-off between automation and transparency, they are said to affect one another because "automation in algorithmic processes could inherently increase opacity." In other words, the more automated that an AI system is, the opaquer it is. In addition, Goodman and Flaxman (2016) and Buiten (2019) refer to the "trade off" between explainability and capacity.

The trade-off between transparency and performance is not limited to the AI literature. For example, in a large literature study Cucciniello et al. (2017) found that transparency can have an effect on government performance. Six studies were found to have a positive effect, one to have a negative effect, five studies with a mixed effect, and one study with no effect (Cucciniello et al., 2017).

2.6.8 Perverse effects

Many examples of perverse effects of AI transparency exist in the literature. The three most prominent examples are regarding: gaming the system, stigmatization, and information bombardment. *Gaming the system* refers to external parties being able to manipulate or evade a system once it is made transparent (de Laat, 2018; Zarsky, 2013, 2016). *Stigmatization* refers to wrongful conclusions regarding certain individuals or groups that are drawn because algorithms and data are made transparent (de Laat, 2018; Zarsky, 2013, 2016). The effect of *information bombardment* refers to the impaired ability for persons to make a decision because they are bombarded with information (e.g. see Buiten, 2019). In this paper, perverse effects are interpreted broadly as the unintended effects that happen as a result of making AI transparent.

2.7 Towards a conceptual framework

The above sections reviewed the theory and observations from existing literature on (AI) transparency to identify the factors, levels, and effects of AI transparency. Not only the concepts were discussed, but also their relationship to AI transparency. The presence or absence of the AI transparency factors

(interpretability, explainability, traceability, accountability, communication) can be considered as the enablers of a certain level of AI transparency (black-box, limited transparency, full transparency). In turn, the level of AI transparency may influence the presence or absence of a variety of effects (accountability, legitimacy, trust, fear, fairness, privacy, trade-offs, and perverse effects). These identified relationships are summated into a conceptual framework (see *figure 3*). This framework will underpin the subsequent chapters of this research.

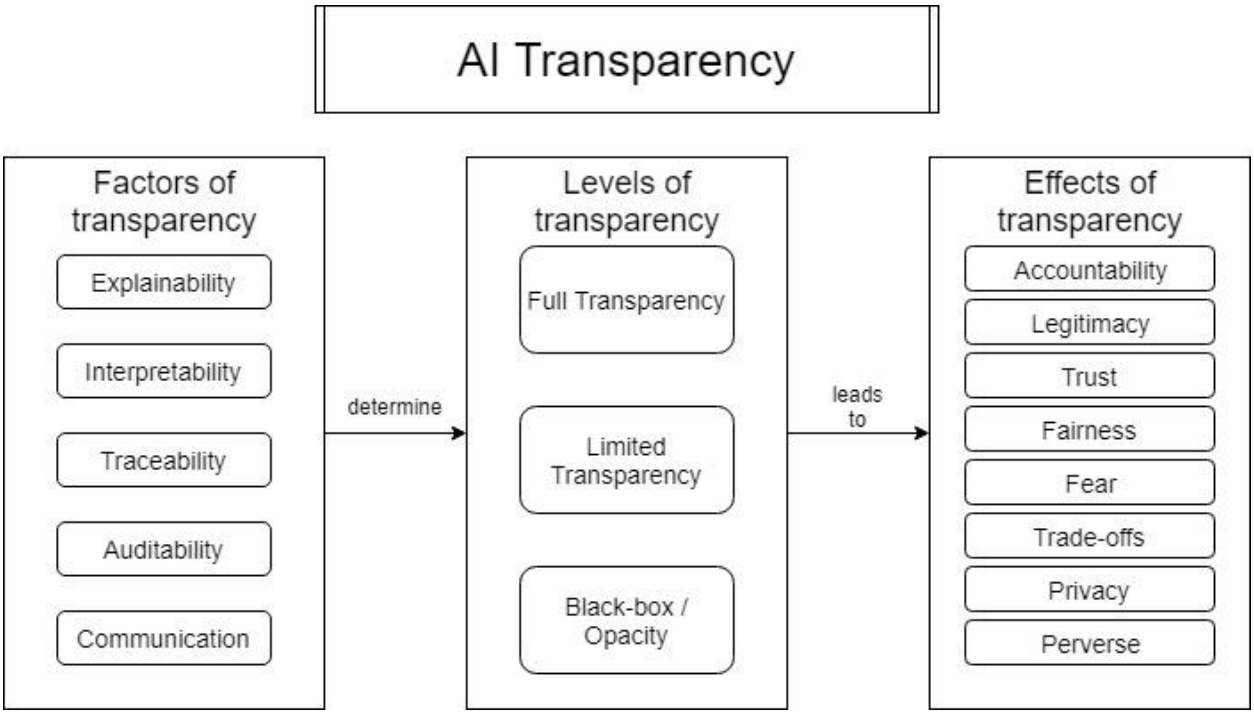


Figure 3. Conceptual framework illustrating the relationship between factors of transparency, levels of transparency, and effects of transparency.

2.8 AI discourses

The purpose of this section is to hypothesize which discourses could be expected for AI transparency. To identify discourse categories, the first step is to briefly discuss what discourses are. The second step is to explore the literature on AI and computer transparency to identify which dominant discourses are likely to exist. The findings of the literature review will culminate into three hypothesized discourses for AI transparency.

2.8.1 What are discourses?

A variety of discourse definitions exist in the literature. A good overview can be found in Gasper and Apthorpe (1996, p. 2-4), some definitions of a discourse exemplified within are: 1) “an ensemble of ideas, concepts and categories through which meaning is given to phenomena;” 2) “any piece of language longer than the individual sentence;” 3) “conversation, debate, [and] exchange;” 4) “an interwoven set of languages and practices” and 5) “a modernist regime [order] of knowledge and disciplinary power.”

This study uses the definition of discourses as used by Hajer and Versteeg (2005, p. 175): a "[d]iscourse' is an ensemble of ideas, concepts and categories through which meaning is given to social and political phenomena, and which is produced and reproduced through an identifiable set of practices."

This definition is used because, as discovered, to *ensemble* a conceptual framework for AI transparency (*the social and political phenomena*) it was necessary to *categorize* (factors, levels, and effects) a variety of related *concepts* (e.g. explainability and interpretability). This study expects that persons will have varying *ideas* regarding the concepts that are part of the AI transparency conceptual framework. Their ideas are expected to influence their perceived *meaning* of AI transparency. The overall meaning of AI transparency for an individual in this study is *a discourse*.

To explain this more concretely. There are two cases which apply here, the differences between which concepts are most important, and the differences between how the same concepts are interpreted. For example, for person A, interpretability and full transparency may matter most. And, person A may emphasize trust because transparency leads to trust. To person B, explainability and limited transparency may matter most. And, person B may emphasize trust because transparency could lead to distrust. The difference between A and B can be interpreted as that the two people have different ideas regarding a similar concept of AI transparency (trust), both in terms of how important the concept is, and how the concept is interpreted. The ideas regarding the concepts in this case influence the meaning one assigns to the phenomena (AI transparency).

2.8.2 Empirical discourses

This study seeks to discover the different produced meanings (discourses) of persons on AI transparency using Q-methodology. The Q-methodology will be discussed in greater detail in *chapter 3*, but a few words are necessary here. The Q-methodology will use the conceptual framework in an attempt to capture the breadth and diversity of the concepts surrounding AI transparency in the form of statements. Participants will then be able to rank these statements against one another and will have the opportunity to comment on statement as well. The commenting of statement will permit to identify how a concept is interpreted, whereas the ranking will allow to identify how salient a concept is. This will enable two discourses to place high salience on the same concept. The comments will then reveal whether high salience was given for the same reason. In short, the ranking of statements as well as the comments given on statements combined result in an empirically measured discourse.

2.8.3 Hypothetical discourses

As mentioned in the introduction of this thesis, the literature on AI transparency discourses and AI transparency public opinion is rather barren. Therefore, this paper investigated the literature on AI and computer mediated transparency to identify which arguments and cognitive structures are likely to exist on AI transparency. Based on the arguments from the literature, AI discourses will be hypothesized. In the analysis chapter (*chapter 5*) this study will investigate how the results match the expectations, and it will report on new insights. These new insights can be used for further academic inquiry.

The first divide that was found is that there are those who favor and others who disfavor transparency. Meijer (2009) investigated of **proponents** and **opponents** of computer-mediated transparency. The proponents find that transparency improves performance (of public officials), it enhances accountability,

prevents corruption, it provides more democracy and affluence, and that perverse effects can be avoided with proper implementation (Meijer, 2009). The proponents also see disadvantages, such as a reduced privacy (Meijer, 2009). De Laat (2018) also highlights arguments that are in favor for transparency (proponents), such as: accountability, the prevention of unjust decisions and biases (including discrimination), to respect the right to know of privacy, and fairness.

The opponents are doubtful whether transparency would provide democratic and affluent gains to society (Meijer, 2009). Opponents worry about the perverse effects of transparency, that it would bring increased uncertainty, that information would blend with misinformation, and that trust would erode (Meijer, 2009). De Laat (2018) also highlights four counterarguments (opponents) to transparency, these are: 1) impairment of privacy through data leakage; 2) perverse effects such as gaming the system; 3) losing a competitive edge; and 4) that AI systems are inherently opaque, transparency would not improve insights.

Beyond proponents and opponents, a third class of arguments regarding AI transparency can be found in Buiten (2019). These are arguments that state that the need for transparency should be "**context-dependent**" (Buiten, 2019). The author explains that the need for transparency should be judged on a case by case basis, depending on: the "risks to safety, fairness, and privacy" (Wachter et al., 2018, p. 4). The author provides several examples, such as: 1) the weighing of trade-offs between transparency and performance for important decisions; 2) to evaluate whether the added value weighs up to the costs; 3) to consider the need for secrecy (e.g. trade secrets); 4) to consider costs and impacts on innovation; 5) to consider the impact on people, where decisions with high impacts may require more transparency; and 6) to consider whether technical explanations are useful for the user. Context-based transparency can also apply within a specific case (de Laat, 2018). One example could be that transparency through explainability is provided to users, but that the transparency of the algorithms is only granted to oversight bodies.

Weller (2017) discusses the benefits and dangers of transparency in intelligent systems. In his paper, the benefits vary depending on the needs of the individual, whether the person is a user, a developer or deployer of the system (Weller, 2017). In sum, the benefits are: understanding how the system works (developer), trust (user), the what and the why of a system (user), understanding the strengths and limitations (society), overcoming "fear of the unknown" (society), understandability of decisions (user), the ability to check and challenge the system (user), auditability (expert), traceability (expert), accountability (expert), legal liability (expert), monitoring and testing (expert), providing user comfort (deployer), and influencing user action and behavior (deployer) (Weller, 2017). The author further emphasizes interpretability and explainability, and that transparency can provide fairness, causality, and verification (Weller, 2017). The dangers of transparency are: manipulation, serving of the deployer's goals, unfair decisions, gaming of the system, privacy (leakage of personal data), trade-offs (such as stifling of innovation and reduced safety), and trust (harsh truths) (Weller, 2017).

It is possible that proponents view AI transparency through a lens of added benefits. Whereas opponents on the other hand might view AI transparency through its possible dangers. And those who analyze the situation from a context-dependent perspective might weigh the benefits and the dangers to come to a decision. In this case, and as mentioned above (see the definition of a discourse), the same related AI concepts can have a different meaning depending on the discourse of a person. For example, for a

proponent, AI transparency leads to trust because they can understand why a system performs certain actions (Weller, 2017). For an opponent, AI transparency may lead to distrust because harsh truths are revealed, or because of information bombardment (Weller, 2017). Whereas for a context-dependent person, in cases where a machine has a low impact transparency may not be needed because trust would not be impaired; in a case with greater impact transparency could be needed because it could impair trust.

Based on the above findings, three discourses for AI transparency are hypothesized:

Hypothetical discourse 1: *the proponents discourse.* The so called "proponents" are those who mainly focus on the benefits of AI transparency. They are also expected to worry about the downsides of non-transparency. Contrary to discourse 2, they would want to see AI to be as transparent as possible. Transparency is expected to alleviate their worries, such as reduced discrimination and the impact on democracy.

Hypothetical discourse 2: *the opponents discourse.* The so-called "opponents" are those who largely worry about the negative impacts that AI transparency might have. They are expected to worry about the potential negative impacts it transparency might have such as impaired competitiveness and innovation, loss of performance, impaired privacy, or the risks for manipulation by external users. They are also expected to reflect on the benefits of non-transparency. Contrary to discourse 1, they are doubtful whether transparency would provide added benefits to society.

Hypothetical discourse 3: *the context-dependent discourse.* The need for AI transparency is "context-dependent" for this discourse. They are expected to carefully consider the context of a case to determine whether the benefits of transparency outweigh the costs of transparency. The context could mean considering trade-offs, assuring fairness, and impact to society. The context could influence what should be transparent, when it should be transparent, and to who. This could mean a preference for limited transparency if the case requires it.

Chapter 3: Discourse analysis using the Q-methodology

This chapter discusses the Q-methodology which is used in this study to empirically measure the discourses of Dutch persons on AI transparency. The first section will briefly introduce the origin and use-case of the Q-methodology to conduct discourse analysis. The second section will cover the process of searching, selecting and classifying the library of statements (the concourse) for this study. To do so, an analytical framework is used (developed on the basis of the conceptual framework from chapter 2). The third section will explain how the concourse was reduced to a set of 54 statements (also called the Q-set) based on predefined selection criteria. The fourth section will explain how participants were selected for this study (the P-set). The fifth section will discuss how the sorting of the Q-set by paticipants was carried out (also called the Q-sort). The sixth section will set the stage for the statistical analysis of the Q-sorts. In the final section, there are a few remarks on the two pilot rounds which were held to prepare for this Q-study.

3.1 Introducing the Q-methodology, a technique to perform discourse analysis

William Stephenson first introduced the Q-methodology on 30 June 1935 in a letter to *Nature* (Stephenson, 1953, p. 8). Stephenson and many academics, notably Dr. Steven R. Brown from Kent State University, have since advanced the methodology. The Q-methodology was "designed to assist in the orderly examination of human subjectivity" (Brown 1980, p. 5). The method can map "how individuals think about an event, issue or topic of research… [and] provide a deeper understanding of the opinions, beliefs, perspectives, decision structures, frames, or narratives of individuals on any topic that has a subjective component" (Brown, Durning, & Selden, 2008, p. 722).

What makes the Q-methodology stand out is that it can capture the "internal standpoint" of persons rather than validating (or invalidating) the "external standpoint" of the researcher, which is often the case in questionnaires and scaling methods (Brown, 1980, p. 1). The Q-methodology can register this internal standpoint through *operant subjectivity* (Brown, 1980). *Operant*, because the participant must weigh and sort a limited set of statements against each other (a process which is called the 'Q-sort'). And *subjective*, because the outcome of sorting the statements are representative of participants viewpoints; these viewpoints can neither be right or wrong because the Q-methodology does not impose *a priori* meanings nor test expectations (Brown, 1980, p. 6; Coogan & Herrington, 2011). What this means for this research is that while a certain set of discourses are hypothesized, there is no necessity that participants will demonstrate one of the expected discourses. Not fitting within an expected discourse can therefore by default not be "wrong."

It is important to note here that subjectivity must be captured by providing a set of statements which represent the main (including complex and diverse) perspectives on the topic (Brown et al., 2008). To ensure that subjectivity is captured (and not my external standpoint), the selection of statements for this study was not on the basis of the hypothetical discourses. Predefined selection criteria were used, which will be discussed in greater depth below.

This study used factor analysis to identify groups of participants who arranged their Q-sorts in similar pattern (Coogan & Herrington, 2011). The totality of this process allowed the discovery of discourses (factors) that are most dominant among the participants (Hermelingmeier & Nicholas, 2017).

The Q methodology is regularly used to conduct discourse analysis. A discourse analysis is a "name of a family of academic research methods that examine, describe and analyze texts… in order to find underlying patterns or meanings" (Webler, Danielson & Tuler, 2009, p. 43). The Q-methodology is used by scholars from a broad range of academic disciplines to explore and identify the discourses of persons on a number of topics (Frantzi, Carter, & Lovett, 2009). These disciplines include political science (Brown, 1980) and public administration (Brown et al., 2008).

Some examples of Q-methodology studies are: Hermelingmeier and Nicholas (2017) who investigated discourses on ecosystem services among ecosystem service researchers. Niedziałkowski, Komar, Pietrzyk-Kaszyńska, Olszańska, and Grodzińska-Jurczak (2018), who looked into discourses on public participation in governance of protected areas. Frantzi et al. (2009) who explored discourses on international environmental regime effectiveness. And co-students at Leiden University de Kleer (2019) and Sluis (2018) who mapped discourses on the freedom of movement of persons. Several primers exist for the Q-methodology, these include Brown (1993), Brown et al. (2008), Newman and Ramlo (2015), Rhoads (2014), Thomas and Watson (2002) and Webler et al. (2009). This study often refers to Webler et al. (2009) as a guiding primer for the Q-methodology. Other primers are consulted on a need-basis as well.

The following sections will discuss the specific steps which must be taken to carry out the Q-methodology. The section for each step will first outline the general guidelines as mentioned in the literature. This will be followed by an explanation of how the step was carried out in this thesis. A brief overview of the steps can be found in *figure 4.*

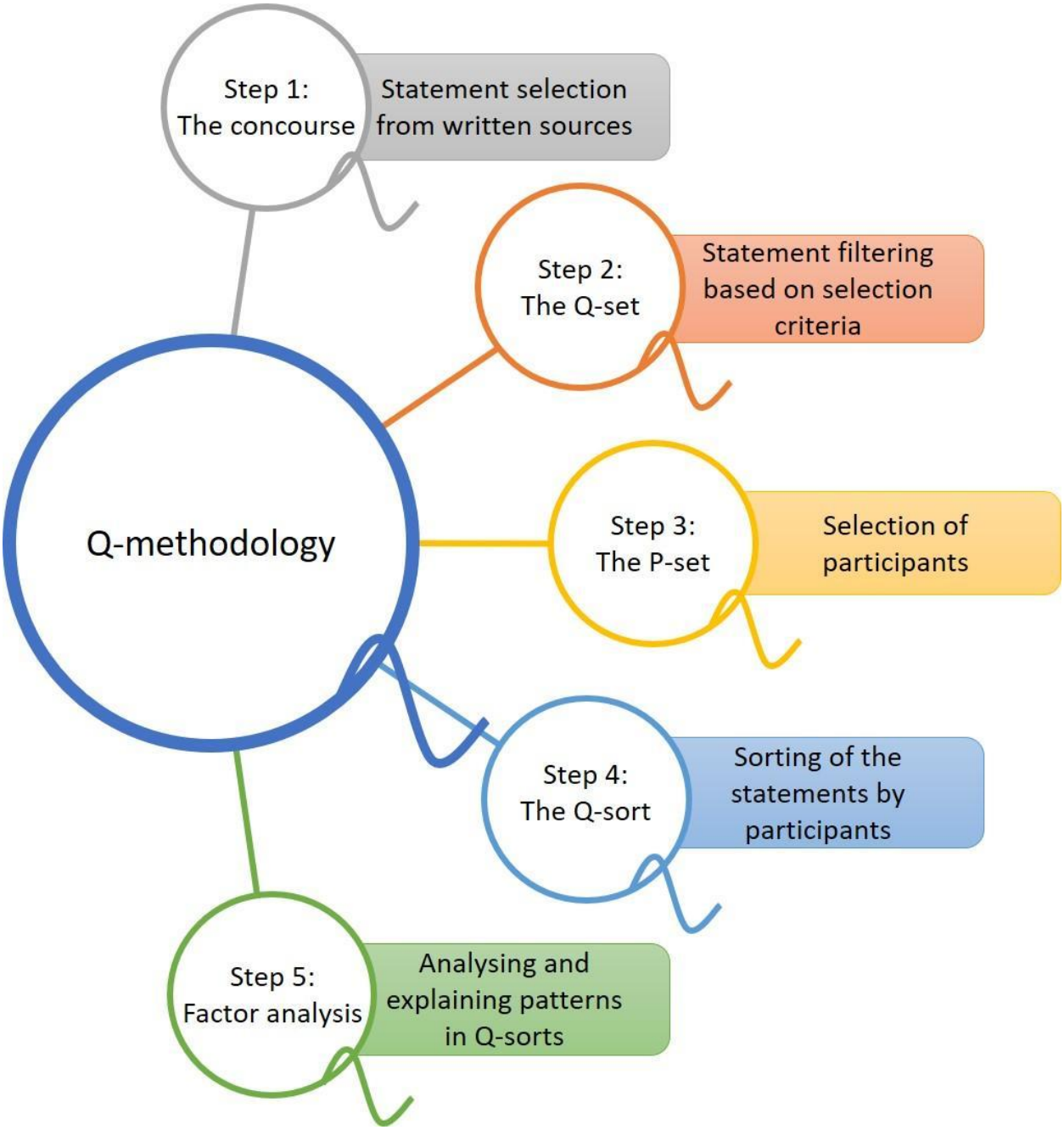


Figure 4: Overview of the Q-methodology steps used in this study.

3.2 Step 1: The concourse

Developing a 'concourse' is the first step of the Q-methodology in this study. The concourse is a large collection of "dozens, hundreds to thousands of expressions of opinions, assertions, and arguments" that exist on a given subject (Brown et al., 2008). The concourse is complete once the totality of expressions that exist on the topic are collected, this is often done in statement-form (Brown et al., 2008). However, Rhoads (2014) mentions that it is likely not possible to capture the totality of expressions which exist on a given topic. Therefore, it is suggested to collect as many as possible of viewpoints on the subject and its

main sub-domains (Coogan & Herrington, 2011b; Rhoads, 2014). The expressions can be collected through interviews, focus groups, from the literature, newspapers, books, blogs, and other written sources (Brown et al., 2008; Newman & Ramlo, 2015; Rhoads, 2014).

To form a concourse for this study, I collected the viewpoints from a diverse set of Dutch written sources. The sources were identified through keyword search, comprising of the keywords derived from the conceptual framework. The keywords used were either first looked up or they were directly translated to their Dutch equivalent (Appendix 1). Not all Dutch equivalent keywords were known at the start of the search, but this changed once more sources were uncovered. Furthermore, some sources used English keywords to refer to a concept, for example, some sources mentioned "artificial intelligence" rather than "kunstmatige intelligentie." Therefore, in some cases English keywords were used to conduct the search. Nevertheless, exclusively Dutch sources were used because the Dutch discourses on the subject are investigated.

To satisfy the need to collect as many as possible viewpoints (Coogan & Herrington, 2011; Rhoads, 2014), statements were collected from consortiums/platforms, news platforms/forums, online newspapers, online media, research institutes/think tanks, branches of the Dutch public sector, political parties, universities, and the private sector (Appendix 2). A wide range of grey literature was used to capture the variety of sources which the participants might read. To diversify search results, the search was conducted through LexisNexis, Factiva, DuckDuckGo, and Google. At the end of the search, 233 statements were collected in the concourse.

3.3 Step 2: The Q-set

Once a concourse is successfully established, a sample named the 'Q-set' must be drawn from the totality of available statements. The Q-set is usually not formed through random selection of statements (unstructured), but rather through a predefined framework (structured) which ensures that the complexity and diversity of viewpoints remain captured (Brown et al., 2008; Rhoads, 2014). Different techniques exist to give shape to this framework, some studies (Frantzi et al., 2009; de Kleer, 2019) use pre-defined frameworks such as the one developed by (Dryzek & Berejikian, 1993). Other studies create their own framework based on sub-categorization of the topic based on findings from the literature, interviews, or other sources (Hermelingmeier & Nicholas, 2017; Kenward, 2019; Rhoads, 2014; Shemmings & Ellingsen, 2014; and Webler et al., 2009). A typical amount of statements used in a Q-set can vary from 30 to 60 statements (Brown et al., 2008; Rhoads, 2014).

3.3.1 Classifying statements

To form the *Q-set* in this study, strategic sampling as mentioned in Webler et al. (2009) was used:

> strategic sampling... simply means that the concourse is divided into categories and the potential Q statements are sorted into these categories. These categories can be theoretically inspired, or they can emerge inductively from a formal or informal analysis. The final set of Q statements… is selected by choosing a small number of statements from each category. (p. 10)

The classification of the statements in this study was based on a self-created analytical framework (*table 1*). The content of the analytical framework was based on the conceptual framework of *chapter 2*. The design of the analytical framework is based on a framework used by Rhoads (2014).

Table 1: Analytical framework used for the classification and selection of statements for the Q-set. Content: adapted from the conceptual framework (*chapter 2*). Design: based on Rhoads (2014).

Category	**Sub-category**	
1. Factors of transparency	a. Explainability	b. Interpretability
	c. Traceability	d. Auditability
	e. Communication	
2. Levels of transparency	f. Full transparency	g. Limited transparency
	h. Black-box	
3. Effects of transparency	i. Accountability	j. Legitimacy
	k. Trust	l. Fear: impact
	m. Fear: risk/safety	n. Fairness: bias
	o. Fairness: competitiveness	p. Privacy
	q. Trade-offs	r. Perverse

3.3.2 Selecting the Q-set

The strategic sampling from the analytical framework was done according to four recommendations from the literature. The first criterion was to select statements that would capture the complexity and diversity of the different views for each sub-category (Brown et al., 2008, p. 723; Webler et al., 2009, p. 18). With the aim to maintain the complexity and diversity of views, subsequent selection criteria were carried out in combination with the first criterion. The second criterion was to narrow further down to the best statements available (Webler et al., 2009, p. 10). This helped to choose the best statement when two (or more) statements were alike. The third criterion was to select statements that would facilitate positive or

negative responses from participants (Webler et al., 2009, p. 18). The final selection criterion was to select an even number of statements from each sub-category of the framework (Brown, 1993, p. 100).

Three statements were selected per sub-category. The selection of statements was not on the basis of the three hypothesized discourses from *chapter 2*. As mentioned above, the aim of the Q-methodology is to capture the internal standpoint of the participant, not the external standpoint of the researcher. Selecting statements based on my external viewpoint could make the outcome of this study flawed.

The initial aim was to have three statements for each of the 16 sub-categories of the conceptual framework. This would result in a Q-set of 48 statements. However, two sub-categories ("fear" and "fairness") were found to have succinct themes within. These themes were identified in the literature as well. To satisfy the need to capture the diversity and complexity of the viewpoints (Brown et al., 2008; Rhoads, 2014), the decision was made to expand these two sub-categories into four sub-categories. The sub-category of "fear" was divided into "fear - impact" and "fear - risks/security." The sub-category of "fairness" was divided into "fairness - competitiveness" and "fairness - bias." The final Q-set selection grid comprised of 18 sub-categories (*table 1*). The selection of three statements per sub-category resulted in a Q-set of 54 statements (see *figure 5*). The complete Q-set of 54 statements is listed in Appendix 3.

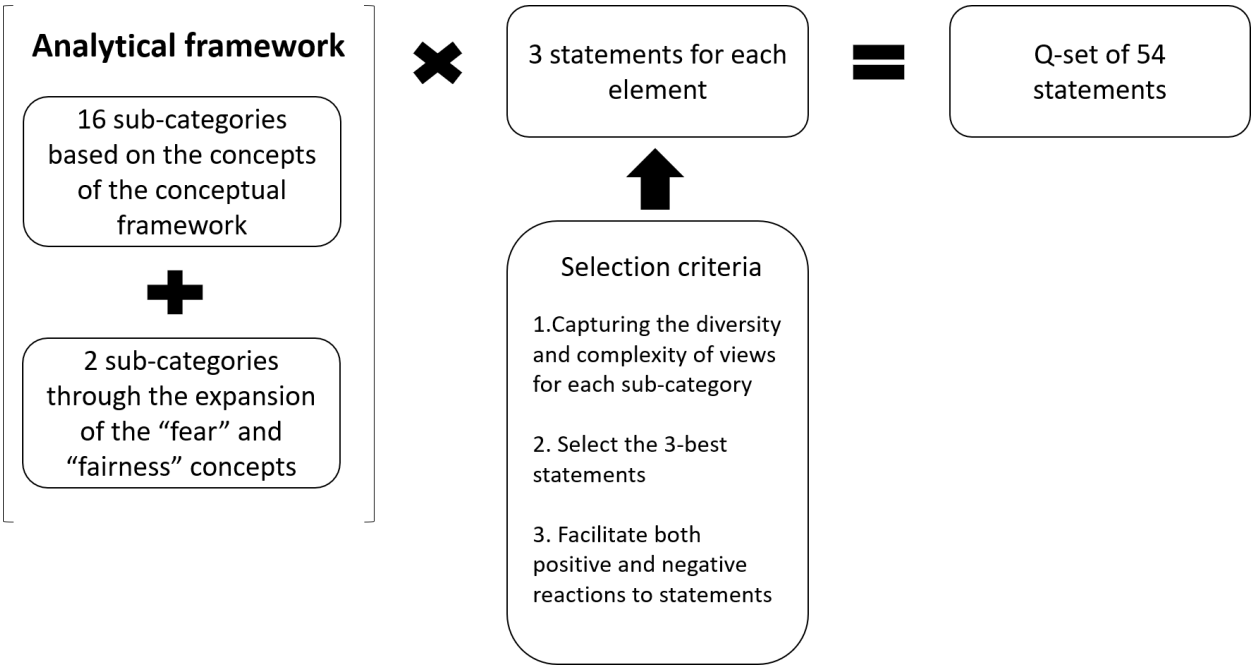


Figure 5: Overview of the selection process for a Q-set of 54 statements.

3.4 Step 3: The P-set

Once the Q-set has been created, a set of participants must be recruited who can engage in the Q-sorting exercise (see step 4 for more information regarding the Q-sorting exercise). This is referred to as the 'P-set', short for persons set. The selection of the P-set does not need to ensure that it is generalizable to the larger population like in a survey study (Shemmings & Ellingsen, 2014; Webler et al., 2009). In the Q methodology approach, the selection of participants should aim that the breath of "positions and opinions

are represented in the P sample" (Brown et al., 2008, p. 723). The aim is not, however, to declare that *all* existing viewpoints are captured, but rather that a diversity of subjective viewpoints can be uncovered (Rhoads, 2014). It is important that the participants selected are relevant to the topic investigated; this can vary to a very broad to a narrow range of suitable candidates. (Rhoads, 2014). The number of selected participants vary by study. Barry and Proops (1999) mention that "as few as 12 participants can generate statistically meaningful results, in terms of the range of implicit discourses uncovered. The reason for this is that each participant's Q sort provides a very large amount of information." Rhoads (2014) mentions that the number of participants can be as many as 40. This range is in line with Webler et al. (2009) who states that "one to three dozen people are sufficient for a Q study."

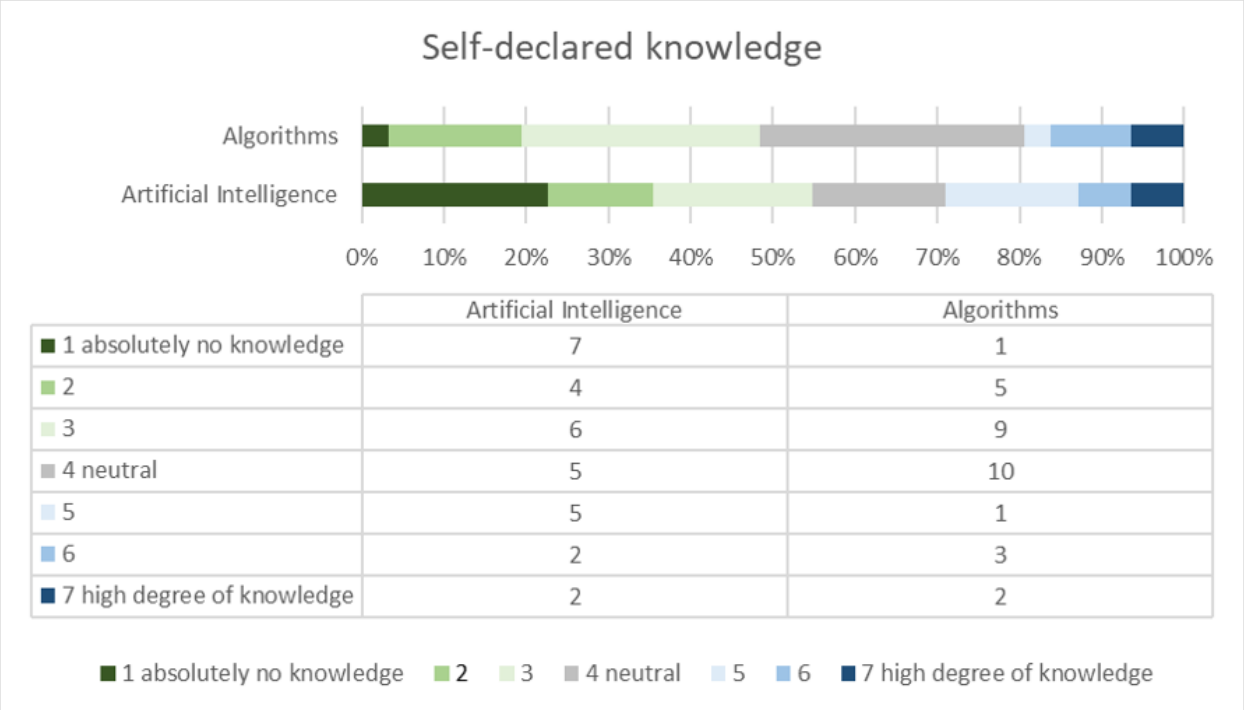
For the recruitment of participants, I based my decision on guidelines provided by Webler et al. (2009), who mentions that to produce a perspective (a discourse) at least three persons must load heavily on the perspective. While my expectations are that three discourses would be unveiled, Webler et al. (2009) mentions that it is "impossible to know ahead of time how many perspectives there are in a concourse, but studies usually produce between two and five." To account for the possibility of five perspectives, this would mean that at least 15 (3 loadings * 5 factors = 15 participants) participants are needed. However, recruiting only 15 participants is problematic because there is a possibility that they will not equally load on each perspective. For example, 5 participants could load on perspective 1, and only 1 participant on perspective 2, and so forth. This would result in failing to register a perspective which in reality might exist. To be certain that up to 5 perspectives can be revealed, the aim of this study was to recruit around 30 participants. This choice was made as a failsafe against the possible heavy loading on select perspectives. As mentioned in Rhoads (2014), the goal is to reveal perspectives and not to claim that all perspectives are discovered. This study recruited 33 participants. The Q-sorts of 2 participants were discarded as "bad data" following the recommendations of Webler (2009, p. 25). These Q-sorts were discarded after the interviews because these candidates were in a great hurry which made it suspect whether their Q-sorts represented their genuine subjective point of view. This resulted in a final P-set of 31 candidates.

To determine the composition of the P-set, this study follows the recommendation by Brown et al. (2008, p. 723) that "[i]f the study addresses a broader topic affecting a larger group of people and interests, the selection of participants should be designed to make sure that the full range of opinions and positions are represented in the P-sample." As discussed in the previous chapter, the subject of this study arguably affects a large group of people in the Netherlands. To ensure that a diverse range of positions and opinions are captured, the following demographic criteria were taken into account: self-declared knowledge of AI and algorithms, profession, and level of education. Gender balance was also sought after, but this was not a primary criterion.

3.4.1 Self-declared knowledge of artificial intelligence and algorithms

The self-declared knowledge of AI (and algorithms) were taken into account because this was also emphasized in two survey studies on the Dutch public opinion of AI (Verhue & Mol, 2018; Araujo et al., 2019). The level of self-reported knowledge was measured during the interview by asking the interviewee to rate his/her knowledge on AI on a scale of 1 to 7 (see Appendix 8). Participants with higher levels of self-reported knowledge does not imply that their opinions will associate with one perspective, because people have selective information seeking behaviour and policy preferences (Nabi, 2003). Nevertheless,

participants with higher self-reported knowledge are expected to have strong opinions on both sides of the issue (Shreck & Vedlitz, 2016). *Figure 6* shows the self-declared knowledge of participants. The self-reported knowledge is comparable to Araujo et al. (2019) where the majority of the participants were below the neutral point (4) of self-reported knowledge for both AI and algorithms. There were more participants with a high degree of knowledge than in Araujo et al. (2019). This was because I actively sought to recruit both knowledgeable and non-knowledgeable participants. Araujo et al. (2019) on the other hand was a large-n survey. An overview of the distribution of self-declared AI and algorithms knowledge is shown in *figure 6.*



	Artificial Intelligence	Algorithms
1 absolutely no knowledge	7	1
2	4	5
3	6	9
4 neutral	5	10
5	5	1
6	2	3
7 high degree of knowledge	2	2

Figure 6: Self-declared knowledge by participants on algorithms and AI.

3.4.2 Profession

Verhue and Mol (2018) also looked into the profession of participants, but to a limited extend as they only targeted civilians and business entrepreneurs. This study will seek to balance the profession of the participants into three broad groups: the public sector, the private sector, and the non-profit/research/academic sector. Participants were also able to declare whether or not they were still a student (see Appendix 8). As discussed in *chapter 2*, profession might influence the opinion of a participant regarding AI transparency. For example, a participant from the private sector could be more concerned about losing competitiveness to other companies when the algorithms of AI are made transparent (Lepri et al., 2018). A participant who is employed in the public sector might be more concerned about the manipulation of government systems when algorithms are revealed (de Laat, 2018). Participants from the non-profit/research/academic sector may again be driven by other motives. An overview of the professional distribution of the participants is shown in *figure 7*.

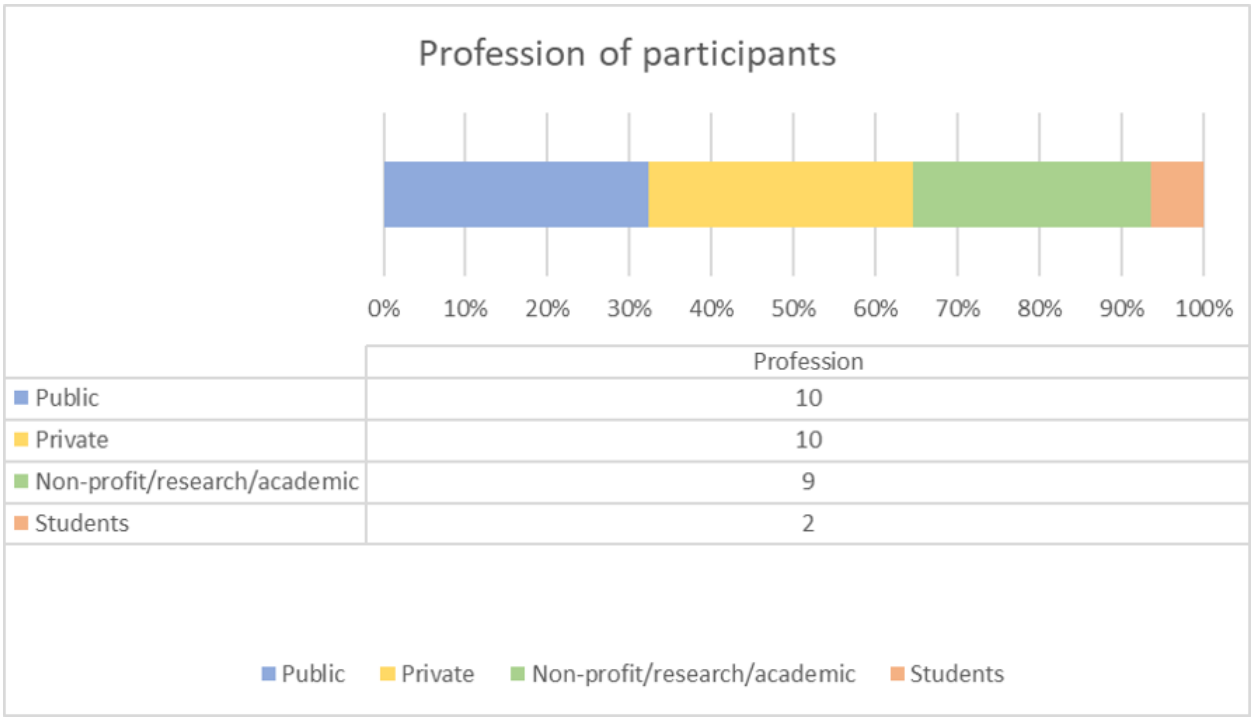


Figure 7: Participants' profession divided by sectors.

3.4.3 Level of education

The level of education of participants were also measured in Araujo et al. (2019) and Verhue and Mol (2018). In this study, participants with various levels of education were therefore also interviewed. This was done to ensure there was no bias would permeate through the overrepresentation of a specific level of education. *Figure 8* provides a summary of participants' level of education.

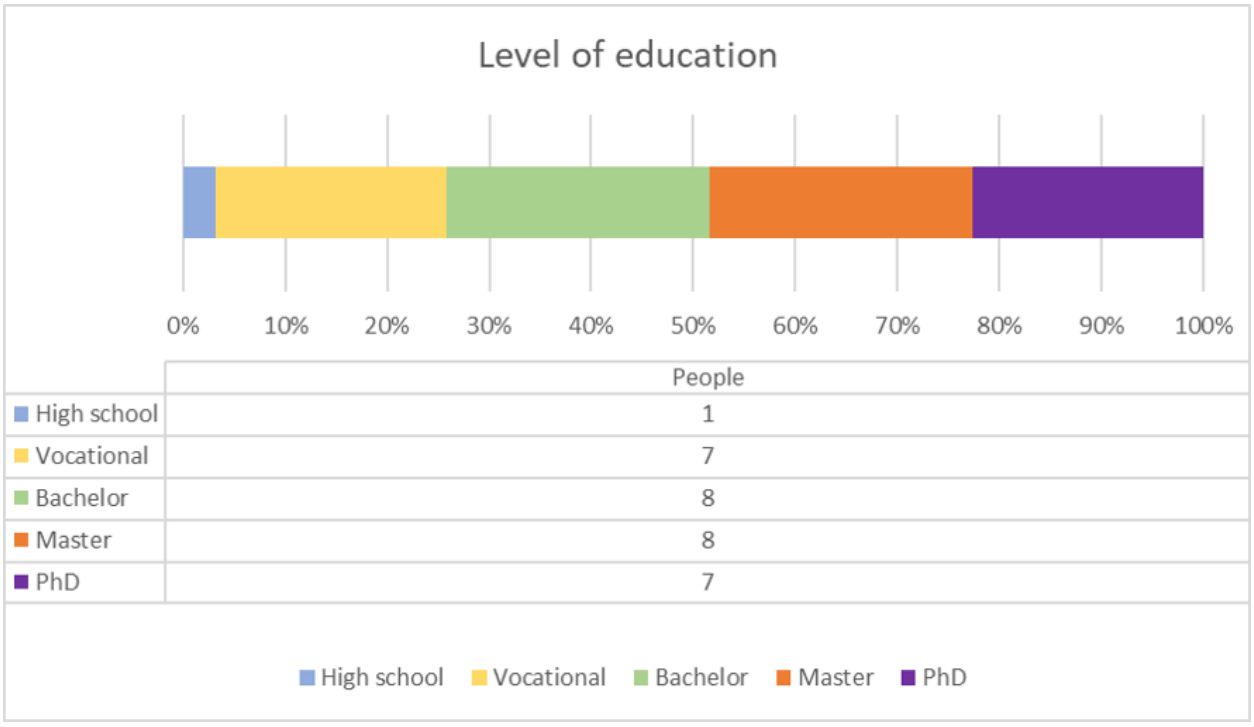


Figure 8: Level of education of Q-sort participants.

3.4.5 Gender and age

This study sought to balance the gender of participants as much as possible, but this was not a primary criterion. The study had 12 female (39%) and 19 male (61%) participants. Participants were required to be 18 and older to participate in this study, as is the case in Araujo et al. (2019) and Verhue and Mol (2018). An overview of the demographic composition of the P-set can be found in Appendix 7.

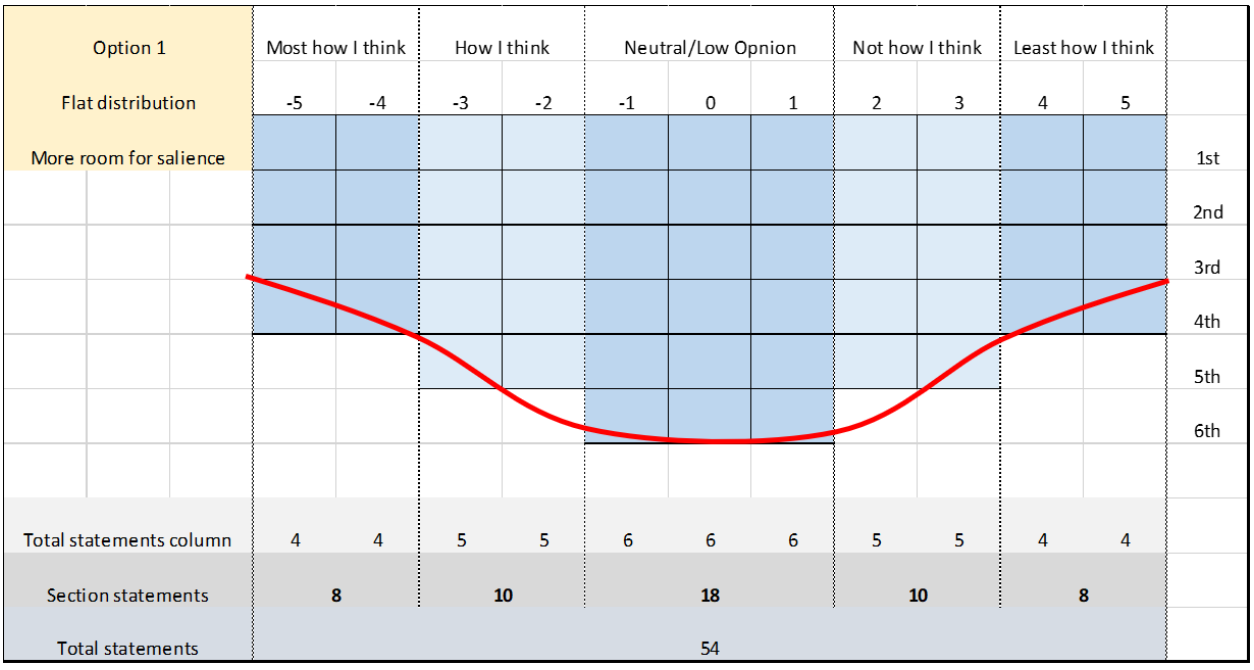
The recruitment of participants was based on snowball sampling as suggested by Webler et al. (2019). Since the objective of the research is to capture different perspectives of the Dutch population, snowball sampling allowed me to reach participants from diverse backgrounds. In addition, when asking for a referral by participants, I was able to communicate which kind of participants I was seeking to select (based on the three primary demographic factors above). However, since snowball sampling is a non-probability sampling technique, the selection of participants is based on the judgement of the researcher (Sharma, 2017). The sample population depends on the choice of participants from the beginning of the research and biases are amplified when the sample size becomes larger (Etikan & Bala, 2017). Nevertheless, I attempt to make a fair recruitment of participants by selecting participants based on the balancing of demographic factors, such as the level of education, profession, and knowledge of AI.

3.5 Step 4: The Q-sort

The Q-sort is a measurement step of the Q-methodology. During the Q-sort, participants are asked to rank the Q-set on a quasi-normal distribution grid (Brown et al., 2008). The quasi-normal distribution grid can be, for example, spread out over a 9-point distribution grid, spanning from +4 to -4 (Rhoads, 2014). Usually the outer ends (e.g. +4, +3 and -3 and -4) are referred to as most agree and most disagree,

whereas the middle values (e.g. +2 through -2) are more neutral stances. Most studies employ a forced distribution method, where participants have a limited number of statements which they can place in any given value (Webler et al., 2009); and some studies employ a free distribution method, where there is no such restriction (Frantzi et al., 2009). One study could not find a statistical difference between studies who use forced distribution or free distribution (Rhoads, 2014). Barry and Proops (1999) recommend using forced distribution when few participants (around 12) and few statements (around 32) are used. During the Q-sort, the comments of the participants regarding the statements may be recorded, and at the end of the Q-sort, the researcher may ask questions regarding the placement of the statements to allow interpretation of the results (Brown et al., 2008; Newman & Ramlo, 2015).

For this study, I used a +5 to -5 quasi-normal distribution grid, where +5 was coded as "most how I think" and -5 coded as "least how I think". The coding according to "how I think" is suggested by Webler et al. (2009, p. 22) who explains that "this leaves open the possibility that the participant could agree (or disagree) with any number of Q-statements." The values -1, 0, and +1 corresponded to neutral opinions. According to van Exel and de Graaf (2005), if the participants are very knowledgeable of the subject, or if a large amount statements are expected to be salient, there should be more room for statements in the strong opinion categories in the outer edges (flat distribution). If this is vice versa, then there should be more room for neutral opinions in the middle ranges (steep distribution). In this study, participants were ranging from very knowledgeable to not very knowledgeable. However, salience was expected over a variety of statements because 18 different AI transparency sub-categories were part of the statements. For example, it could be the case that 10 of the sub-categories would be of salience to the participant. Forcing participants to place some of these statements in the neutral ranges could impact the outcome of this study. Therefore, the decision was made to make the quasi-normal distribution grid flatter. This would allow for the expression of more salience compared to a steep distribution (see *figure 9*).



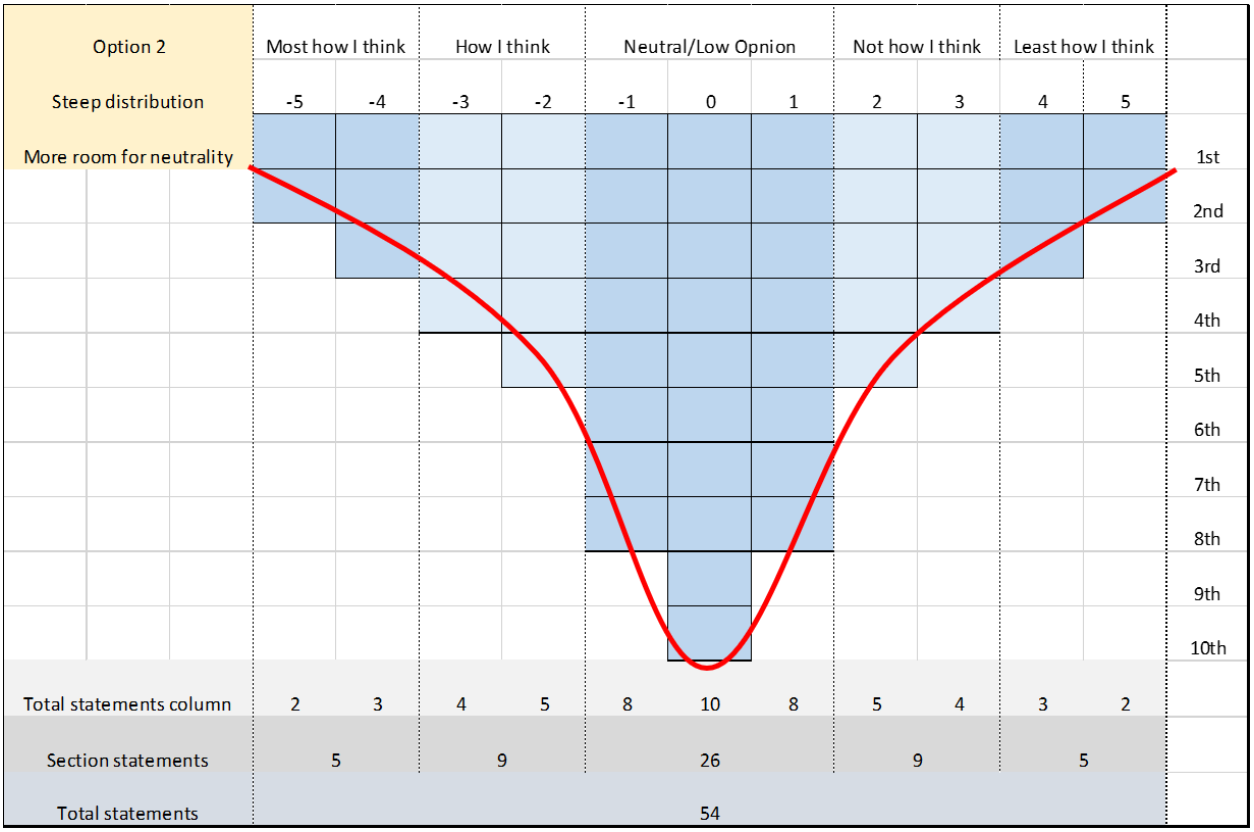


Figure 9: Two options considered for the quasi-normal distribution of the Q-sort. The red line is included to illustrate the shape of the distribution. Option 1 (top) refers to a flattened quasi-normal distribution with more space in the salient categories. Option 2 (bottom) refers to a steep quasi-normal distribution with more space in the neutral categories.

At the beginning, participants received an explanation on how to correctly execute a Q-sort, the research rationale, and the societal and academic relevance of this study. Participants were explained that answers were based on personal opinion (subjectivity), which meant that neither wrong or right answers could be given. The participants were encouraged to make comments during the sorting process as so desired. These comments were recorded and subsequently used for analysis. Each statement card had a random number from 1 to 54. The placement of the statements by the participants was recorded by writing the down the numbers of the statements. The recording of the Q-sort was subsequently verified by the participant. At the start of the Q-sort, participants received a short questionnaire (see Appendix 8). And at the end of the Q-sort, the participants were able to mention where their midpoint was located (the transition point between “how I think” and “not how I think”), to comment on why they chose specific statements, and to give an overall take-away message for their Q-sort (Appendix 8). The midpoint was registered because there is a possibility that participants agree to more statements than disagree with statements (or vice versa), which would mean that their midpoint would no longer be on the 0 value of the distribution grid (Webler et al., 2009). On average, the midpoint of the 31 participants was situated at -1 and not 0. This did not necessitate interpreting the results differently because -1 is still located in the ‘neutral’ zone of the distribution grid.

3.6 Step 5: Factor analysis and factor interpretation

Once the Q-sorting has concluded, the final step of the Q-methodology is to conduct a factor analysis on the Q-sorts to infer results. This study uses the PQMethod software (version 2.35) by Schmolck (2014) to conduct principal component analysis (PCA) and automated varimax rotation for the factor analysis of the Q-sorts. The results of the factor analysis will be thoroughly discussed in the results chapter (*chapter 4*) of this thesis.

Some datasets, as is the case for the many collected Q-sorts, can appear messy at face value. This can be seen in the correlation table of Q-sorts (Appendix 4). There appears to be no real structure in the data. This is where factor analysis has a valuable role. Factor analysis "is a mathematical technique that reveals underlying explanations for patterns in a large set of data" (Webler et al., 2009, p. 26). The 'factors' can be best viewed as a higher level of correlation between groups of Q-sorts which share similarities (Rhoads, 2014). Factors in fact are hypothetical Q-sorts, to which a group of Q-sorts have a certain degree of correlation**.**

Because factors reveal a pattern across a group of Q-sorts, they are sometimes referred to as "idealized sorts" or "social perspectives" (Webler et al., 2009). In principle, for the Q-methodology, data is analyzed through inverted factor analysis, "which means that instead of looking for patterns across people's responses… [it looks] for patterns across Q statements" (Webler et al., 2009, p. 44). This means that the Q-sorts of people are the variables to be analyzed, not the individual responses to specific questions (Webler et al., 2009).

One mathematical method that frequently used to carry out factor analysis for the Q-methodology is called 'principal component analysis' (PCA). PCA identifies which factor can explain as much as possible of the variance in the data. PCA produces n-factors, starting with the factor that captures most of the variance, a second factor that captures the second most variance, and so forth, until all of the variance is captured. Subsequently, varimax is an automated way to rotate the factors, this rotation seeks to capture as much as possible of the variance on a predefined number of factors (Webler et al., 2009).

In this thesis, factors will be referred to as "discourses." This is because, as we have discussed above, it represents a hypothetical Q-sort which comes as close as possible to as many as possible of the Q-sorts of the participants. Correlations are used to express how closely a Q-sort corresponds to discourse (read: factor). These correlations are also referred to as "loadings," where a high loading means a high correlation (e.g. see Brown et al., 2008).

The loading (correlation) of a Q-sort to a discourse has to be in excess of 0.351 for it to be significant at $p < 0.01$. The following formula (Brown, 1980, p. 223) was used to determine statistical significance at $p < 0.01$:

$$f = 2.58\left(1/\sqrt{number\ of\ statements\ in\ a\ q-sort}\right)$$
$$f = 2.58\left(1/\sqrt{54}\right) = 0.351$$

The loading (correlation) of a Q-sort to an ideal sort has to be in excess of 0.267 for it to be significant at $p < 0.05$. The following formula (Brown, 1980, p. 223) was used to determine statistical significance at $p < 0.05$:

$$f = 1.96\left(1/\sqrt{number\ of\ statements\ in\ a\ q-sort}\right)$$
$$f = 1.96\left(1/\sqrt{54}\right) = 0.267$$

One manual component of factor analysis is to determine which Q-sorts are representative of a specific factor, this process is also called "flagging." Flagging matters "because the final description of each factor will be based on a weighted average of only those sorts flagged as loading on that factor" (Webler et al., 2009, p. 31). In the Q-methodology, the Q-sorts are the variables (Webler et al., 2009), and Q-sorts which significantly load on two factors can act as a confounding variable (Watts & Stenner, 2005). To facilitate that two factors would be discernible from one another and to prevent the introduction of confounding variables, the following rules were applied for the flagging of Q-sorts:

1. a Q-sort can be flagged if it has a loading of 0.351 or higher ($p < 0.01$) to one factor
2. a Q-sort cannot be flagged if it has a loading of 0.351 or higher ($p < 0.01$) to two factors
    a. For example, if a Q-sort has a loading of 0.41 to factor 1 and a loading of 0.37 to factor 2, it cannot be flagged to either factors
3. a Q-sort cannot be flagged if it has a loading of 0.351 or higher ($p < 0.01$) to one factor, and a loading of 0.267 or higher ($p < 0.05$) to a second factor
    a. For example, if a person has a loading of 0.58 to factor 1 and a loading of 0.29 to factor 2, it cannot be flagged to factor 1

The flagged Q-sorts can be found in Appendix 5.

3.6.1 Data management and informed consent

At the start of the Q-sort participants were asked to agree to participate in the study by signing a consent form (see Appendix 9). And at the end of the Q-sorting exercise, the information of the participants' questionnaire and registered Q-sort was processed using a random ID-number (for ease of data-processing, random ID-numbers that started with a 0 were not used). This step was taken to prevent that the identities of the participants would be revealed. The identities of the participants were solely on the consent form, in paper form. A few participants received a copy of their consent forms by email, these scans were removed from the computer and the e-mail outbox once their consent forms were sent.

3.7 Two pilot sessions to improve reliability

To ensure consistency throughout the Q-sorting process, two pilot sessions were held. Each pilot session had three candidates. Feedback from the first pilot session facilitated in making the statements more comprehensible. It was mentioned that some statements were too complicated, that they could benefit from some simplification to understand the statements better. And it was mentioned that some statements appeared to have a double meaning, making it difficult to place it on the distribution grid. Based on the feedback, the complicated statements and the statements with a double meaning were simplified. During

the second pilot session, a few additional statements came to the fore which were difficult to interpret. These statements were marked with a *, but they were no longer simplified and reprinted. Subsequent non-pilot participants were informed that the statements marked with a * were more complicated. I suggested to ask for my help if they appeared too difficult to understand.

Chapter 4: Empirical findings

This chapter will report the results of the factor analysis on the Q-sorts of 31 participants. As mentioned in *chapter 3*, the factor analysis was done through principal component analysis (PCA) and varimax rotation, using the PQMethod software (version 2.35) by Peter Schmolck (2014). Section 4.1 explains the number of factors that would represent the data in the most robust manner. Section 4.2 illustrates the flagging rules used to minimize confounding variables. Section 4.3 briefly discusses how factors are interpreted as discourses. Section 4.4 - 4.6 then describes the three discourses individually based on the most salient statements, and enriched with the comments from the flagged participants of the discourse.

4.1 Determining the amount of factors

To determine the number of factors to analyze, this study follows the suggestion that at least 3 participants must load highly on a factor (Webler et al., 2009). For the purpose of robustness, a second criterion that was included was to account for how many of the total participants were highly loaded on only one of the factors. A varimax rotation of two, three, four, and five factors was performed. Following the three flagging rules (described in *chapter 3*), the total amount of high loadings for only one factor dropped noticeably in the 4 factor and 5 factor rotations compared to the 2 factor and 3 factor rotations.

For the 2 factor rotation, all factors are well above the requirement of having 3 high loadings. The lowest high loading included is for factor 2 at 0.364, and the second lowest high loading is also for factor 2 at 0.3911. Out of 31 participants, 21 (68%) are loaded highly on only one of the factors. Because the number of high loadings per factor are well above the requirement of 3 high loadings, a 3 factor rotation is done to examine whether more factors could be found in the data.

For the 3 factor rotation, all factors are still well above the requirement of having 3 high loadings. This illustrates that another factor can be identified in the data. All of the high loadings are well above the 0.351 threshold ($p < 0.01$). The lowest high loading included is for factor 1 at 0.433, and the second lowest high loading is for factor 2 at 0.459. Out of the 31 participants 21 (68%) are highly loaded on only one of the factors. This means that there was no drop in the total amount of high loadings for only one factor compared to a 2 factor rotation.

For the 4 factor rotation, factor 2 barely makes the requirement of having 3 high loadings on one factor (one loading with 0.3676 was just slightly above the threshold of rule 1). Seventeen participants of the total 31 (55%) are loaded highly on only one factor. This is a drop of 4 candidates (13%) compared to a 3 factor rotation.

For the 5 factor rotation, only factor 1 makes the requirement of having 3 high loadings on only one factor. And 13 out of the 31 participants (42%) are loaded highly on one factor. This is a drop of 8 candidates (26%) compared to a 3 factor rotation, and a drop of 4 candidates (13%) compared to a 3 factor rotation.

Based on the number of participants with high loadings on only one factor, and the number of high loadings per factor, this study found that a 3 factor rotation produced the most robust results for this data-set. No drop in the number of high loadings was found between a 2 and 3 factor rotation, but the shift to 3

factors revealed the presence of a third factor with 5 high loadings (*table 2*). A drop in the total number of high loadings was found when the analysis moved from a 3 factor to a 4 and 5 factor rotation. The number of loadings per factor dropped drastically as well 4 and 5 factor rotations, with the 4 factor rotation just barely making the 3 high loading requirement for its factors, and the 5 factor rotation only one factor making the 3 high loading requirement.

Table 2: An overview of the number of high loadings per factor on different factor rotations.

Factor	**Number of high loadings on only one factor**			
	2 factor rotation	**3 factor rotation**	**4 factor rotation**	**5 factor rotation**
Factor 1	11	10	7	7
Factor 2	10	6	3	1
Factor 3		5	4	1
Factor 4			3	2
Factor 5				2
Total high loadings on one factor (%)	21 out of 31 (68%)	21 out of 31 (68%)	17 out of 31 (55%)	13 out of 31 (42%)

4.2 Verification of the flagging rules

Chapter 3 explained that three flagging rules are applied to determine whether participants are loading highly on only one factor. These rules are applied to eliminate confounding variables and to make factors discernible from one another. The following three rules were applied:

1. a Q-sort can be flagged if it has a loading of 0.351 or higher ($p < 0.01$) to one factor;
2. a Q-sort cannot be flagged if it has a loading of 0.351 or higher ($p < 0.01$) to two factors;
3. a Q-sort cannot be flagged if it has a loading of 0.351 or higher ($p < 0.01$) to one factor, and a loading of 0.267 or higher ($p < 0.05$) to a second factor

A trial run was made to verify this statistical assumption, the results are displayed in *table 3*. The correlation between the factor scores decreased with the application of each additional rule. One increase in the correlation between factor 2 and 3 occurred when going from applying 1 rule to 2 rules. This could have occurred through the introduction of a confounding variable. For example, one participant had a loading on factor 2 of 0.4194 ($p < 0.01$) and a loading on factor 3 of 0.3152 ($p < 0.01$). Through the introduction of the third rule, this Q-sort and possible confounding variable was no longer flagged. With the application of all three rules, the correlation between all factors dropped to their lowest levels. Seven

high loadings were not flagged because of the application of the flagging rules. Two high loadings were not flagged because of rule 2, and 5 high loadings were not flagged because of rule 3.

Table 3: Correlation between factor scores: with only 1 rule, with rule 1 and 2 (2 rules), and with all rules (3 rules). The lowest correlation between factors was achieved through the application of all three rules. A lower correlation between factors implies that the factors are more discernible from one another. None of the factors are correlated significantly to another factor (at $p < 0.05$). Red indicates the highest correlation, light green a reduced correlation, and dark green the lowest correlation.

	Factor 1			Factor 2			Factor 3		
	1 rule	2 rules	3 rules	1 rule	2 rules	3 rules	1 rule	2 rules	3 rules
Factor 1	1.000	1.000	1.000	0.188	0.149	0.091	0.161	0.088	-0.007
Factor 2	0.188	0.149	0.091	1.000	1.000	1.000	0.274	0.283	0.258
Factor 3	0.161	0.088	-0.007	0.274	0.283	0.258	1.000	1.000	1.000

4.3 From factors to describing discourses

As mentioned in *chapter 3*, the resulting factors after PCA and varimax rotation are sometimes called "ideal sorts" or "perspectives" (Newman & Ramlo, 2015; Webler et al., 2009) or "points of view" (Rhoads, 2014). This is the case because factors represent a hypothetical Q-sort, which correlates significantly to a number of the Q-sorts of the participants. In this study, the factors will be representative to a *discourse* that a group of the participants were found to have.

To interpret a factor as a discourse, the first step is to analyze the statements that most salient for the factor (Newman & Ramlo, 2015; Webler et al., 2009). The most salient statements are those that are placed in the +5 and -5 column (called factor arrays, see below). Statements for each discourse receive factor arrays based on their Z-score. For example, the forced distribution grid has room for four +5 statements. Therefore, the four statements with the highest Z-scores automatically receive a +5-array score.

Z-scores are useful to identify by how many standard deviations a statement is distanced from the mean of the statements. Z-scores measure the salience of a statement for a discourse. For example, a statement with a Z-score around 0 implies that the participants who highly correlate to this discourse, on average, placed this statement is in the neutral area of the sort. Note that "on average" is written because it could be that one participant would assign a +4 and another participant would assign a -4. On the other hand, the more positive or negative that a Z-score of a statement is, the greater the average salience this statement would have for the participants. For example, for a high positive Z-score, on average more participants would have placed this statement in a +5 category than a statement with a 0 Z-score.

The Z-score can also be used to identify which statement is more salient amongst the statements which have the same factor array (Appendix 6). For example, for factor 2, the highest ranked statement has a Z-score of 2.181, and the second highest ranked statement has a z-score of 1.478. While both statements have a factor array of +5, the difference of a standard deviation of 0.703 illustrates that the highest ranked statement is much more salient. On the contrary, if the difference of the Z-scores is smaller, it means that the two statements are comparable in terms of their salience. Solely reporting a factor array (e.g. +5 or -5) would not illustrate this subtle but potentially important difference between statements.

Once the salient statements for each factor are identified, the researcher can write out an interpretation of the discourse. In the subsequent sections, each discourse is first interpreted based on the salient statements. To decipher why salience was given to a statement, each interpretation is then supplemented with comments from the participants.

4.4 Discourse A: no excuses, we demand AI transparency!

4.4.1 Discourse A interpretation based on statements with high salience (see table 4)
Persons with this discourse are transparency pioneers; transparency is by all and for all. They find transparency a requirement for AI decisions to be legitimate (12). To them, it must be clear who is accountable for an AI decision, transparency can ensure that humans are held responsible and not the algorithm (09). This discourse finds transparency to be an important feature for the functioning of a democratic society, effectiveness can therefore not be a reason to be less transparent (16). They worry that opacity shifts the balance of power between civilians and the state (07). The right kind of transparency is needed to gain the trust of people (08).

Persons who align with discourse A are not willing to accept arguments that are in favor of opacity. They do not agree that AI systems are better than humans precisely because they are opaque (20), or that explainability limits the added value of AI (33). They are unwilling to accept arguments which claim that people should get used to computer decisions being a black-box (32). Other arguments against transparency are not welcomed either, such as that AI transparency is complicated and incomprehensible, and that transparency does not contribute to trust (08).

Table 4: Discourse A high salience statements. Statements with * are discourse distinguishing statements at $p < 0.05$, with ** at $p < 0.01$.

Z-score	+5 (most like how I think)	Z-score	-5 (least like how I think)
1.624 **	12. Transparency can make an important contribution to the legitimate application of AI.	-1.864 *	20. The way that AI systems are self-aware and achieve goals is often not transparent, and that is exactly why they are sometimes better than humans.
1.587 **	07. The secret character of activities in the security domain strengthen the transparency paradox: civilians become more transparent for the government; while the algorithms of the government are barely transparent. This causes the balance of power between the civilians and state to shift.	-1.814 **	33. Explainability can limit the added value of algorithms, because explainability places limits on the complexity that an algorithm can have.
1.577 **	09. Transparency is important to get clarity about accountability; humans are and will remain responsible, and not 'the algorithm'.	-1.805 **	32. We should get used to some computer decisions being a black-box, just like human decisions.
		-1.792	08. Openness about the algorithm is

1.576 **	16. Losing effectiveness may, in a democratic society, not be a reason to drop the obligation for transparency.	**	difficult because users often do not understand how such an algorithm works. Such an explanation is too complicated and does not contribute to trust.

4.4.2 Discourse A participant comments

Participants who aligned with discourse A had an abundance of comments to refute the statements which call for less transparency or opacity. For example, participant 3722 acknowledged then rebuffed statement 20 (see *table 4*), saying that *"this is or can be true, but this does not mean that we must not try to understand it."* Statement 20 was also rejected by participant 6277, who said that *"the concept of opacity cannot contribute to better AI. This would be like creating a noise and then saying that that makes [AI] better."* And on statement 45 (on whether we should limit the visibility of algorithms to prevent that the privacy of individuals can be affected) participant 3722 mentioned that *"privacy and openness are of equal importance, and that it is the task of the AI maker to treat these equally."* Participant 6277 refuted statement 32 (see *table 4*) saying that *"we mustn't want to get used to this. To implement AI, along with all its risks, I believe that we must try to understand it's decisions."* Participant 9916 refuted statement 32 just as strongly, saying: *"I must get used to it, but I rather would not want to. We have no choice, we are neatly walking behind the elites. If it could be possible then I would not want to get used to this."* Participant 9916 acknowledged and rebuffed statement 47 (on impaired innovation as a result of transparency), saying that *"I do understand this argument, but this cannot be a reason to not be transparent."* Participant 5993 refuted statement 15 on limited transparency, saying that *"this is a cutoff way to not be transparent. Either you are transparent, or you are not. A limited form of transparency for me means that you are not honest, and that you do not want to tell the truth."*

The importance of legitimacy and acceptance came to the fore in many comments. For example, participant 3722 said regarding statement 9 (see *table 4*): *"until AI has been declared as a legitimate self-aware being, until then the AI maker is always responsible."* Participant 6277 also highlighted legitimacy in his comment on statement 02 (whether government decisions must be transparent to be legitimate) saying that *"to be legitimate, the civilian must understand what is being done, this can only happen through transparency in a democratic society."* Participant 5993 too commented on statement 02, saying:

> *if the government wants to use algorithms on such a large public audience, then you must be able to justify it. Naturally persons will ask a lot of questions. If you provide an explanation, then you will be convincing and earn trust*

Participant 6277 further commented on statement 32 that "*In the future it will concern more use-cases, we must build on our values and therefore a deficit of transparency shouldn't be accepted."*

Accountability and the centrality of humans was a recurrent remark. Participant 7722 linked traceability to accountability in a response to statement 22 (on whether decisions of robots are traceable), saying that *"it must be traceable to humans. Information and knowledge is power, and if it is not traceable, you will get a power imbalance, that must not happen. We must know who is responsible."* Along similar lines, participant 6277 commented on statement 22 that *"the criteria must be created by humans and then the*

system will be traceable." Participant 7722 rejected statement 20 (see *table 4*), stating that *"AI systems must ideally always be under the control of humans, otherwise we will find ourselves in a dangerous situation."* Human centrality was also in a general comment by participant 9916: "*the ethical dimension is important, the story of humans, humans must be central."* Participant 4610 commented on statement 09 (see *table 4*) that *"humans are and will remain responsible."* Participant 9542 commented on statement 36 (on the limits of technical transparency) that *"someone must understand how it works. I cannot know everything, but there must be humans who understand the algorithm."*

The right approach to transparency was often emphasized as a means to earn the trust of people. Participant 6277 refuted statement 08 (see *table 4*) saying:

> *you should just make everything as simple as possible. If you cannot be open and you cannot give an explanation, then people cannot trust it. The principles and basis must be well explained in order to win trust.*

Participant 3722 commented on statement 41 (which poses whether to not trust a machine if it cannot give a better explanation than humans) that *"it is not for the machine to explain itself, it should be the maker."* And participant 3722 later remarked for statements 5 and 27 (on the need for transparency for legitimacy) that *"without transparency, there cannot be trust."* And, with transparency, a demand for the right approach to transparency is expected to increase trust; for example participant 9916 made the general remark that *"what also isn't transparent is when you receive high volumes of text and then you just have to agree to it all. This does not contribute to trust."* Participant 4610 commented that *"transparency is important because humans trust systems blindly."* Participant 9542 commented on statement 8 that *"you must be able to explain [algorithms] one way or another in order to make humans understand and trust it."*

Comments made it clear that it should be about transparency by everyone for everyone, otherwise there could be an imbalance between groups in society. For example, regarding statement 38 (on whether residents must maintain full control over their personal data) participant 3722 replied that *"'resident' implies that this only concerns the government, but I find this even more important for companies."* And regarding statement 7 (see *table 4*) participant 7722 mentioned that this is *"an enormous problem, because of the difference in power between people and the state you can get a high number of negative outcomes, such as discrimination, exploitation, and use for wrong purposes."* Participant 6277 commented on statement 44 (on more nuance from the government regarding algorithmic supervision) saying: *"I do not align with this opinion, in my opinion the government must protect civilians by ensuring transparency, especially against large commercial companies. There is already too much nuance."* Participant 6277 also commented on statement 54 (on that the choice of the algorithm depends on interpretability and explainability) that *"in algorithmic decision-making, explainability and interpretability are one of the highest priorities. You must understand [the algorithm]."* And participant 9916 commented on statement 37 (regarding that auditing is possible by judges and authorities, but not by civilians) that *"this here means that you will encounter the elites. This ensures that civilians will be kept stupid. From an ethical point of view civilians must get insights."* And even in the case of national security, participant 4610 commented on statement 17 (on exempting transparency requirements for

lawful reasons such as for national security purposes) that: *"I would want to know why that would be a threat to national security. I want an explanation why information should be kept secret."*

4.5 Discourse B: balance the needs for AI transparency!

4.5.1 Discourse B interpretation based on statements with high salience (see table 5)
Persons who identify with this discourse believe that it is not always in their best interest to have AI transparency. They are in high agreeance with reduced transparency as a trade-off for the effective operations of intelligence services (13). Beyond the state, they also believe that we should balance the need for transparency and secrecy for commercial interests (21). These arguments are in line with the general notion that limited transparency should be considered for some cases (15).

Those aligned with this discourse are not appealed to arguments that warn against limited AI transparency. They are still able to trust a machine if it does not give a better explanation than humans (41). They do not believe that most risks of AI systems occur precisely because they are not transparent (03). Moreover, they do not believe that only good AI performances should be publicized, and that those who fail should not (29).

This discourse sees the added benefit of auditing and tracing back to mistakes. In cases of limited transparency, they would agree that some forms of transparency can be a helpful tool, such as the logging and storing of information regarding the decision process that can be used during audits (23). They do not agree with arguments that state that AI decisions cannot be traced back to their origin (22).

Table 5: Discourse B high salience statements. Statements with * are discourse distinguishing statements at $p < 0.05$, with ** at $p < 0.01$.

Z-score	+5 (most how I think)	Z-score	-5 (least how I think)
2.181 **	13. Intelligence services cannot effectively operate when they have to be fully transparent.	-1.849 *	22. It is not traceable whether a robot takes decisions on the basis of valid criteria (incorrect or non-scientific results are not valid).
1.478 **	21. The art is to find a balance between algorithmic transparency on the one hand and the commercial interests of secrecy on the other hand.	-1.670	29. Only research that 'demonstrates' that the robot performs well will be made public. Because of this there will be a faulty reputation about the quality of the robot.
1.474	23. The decision process of machine learning technology must be logged and stored. In such a manner that this information remains available for a minimum period to be inspected during audits.	-1.548 **	03. Many of the risks of AI happen as a result of a limited transparency of AI.

1.432 *	15. Next to full transparency there are some cases where more limited forms of transparency could be considered.	-1.306	41. If the machine cannot give a better explanation than humans, then do not trust the machine.

4.5.2 Discourse B participant comments

Participants with this discourse repeatedly commented on the need to consider the context of a case to determine the need for transparency. Participant 5636 commented on statement 21 (see *table 5*) that *"a balance for the entire topic is highly important. There is a large grey area. I understand the interests of both sides. You do not want to inflict economic damage."* Participant 9279 also commented on statement 21, saying that *"companies and governments must also be able to do their work well to be of added value. A balance is therefore necessary."* On statement 37 (regarding that auditing is possible by judges and authorities, but not by civilians) participant 5636 mentioned that *"there are serious parties needed who receive insight. Giving the civilians insight will only bring misery."* Participant 8711 commented on statement 34 (which states that transparency of parts should depend on specific needs) and 28 (on not allowing algorithms to be a black-box) that *"this is dependent on the situation."* Participant 5215 commented on statement 15 (see *table 5*) that *"[transparency] depends on the application and the requirements, transparency is not a solution for everything."*

Depending on the situation, the need for opacity was agreed upon. For example, participant 8711 commented on statement 24 and statement 10 (both on transparency and algorithm manipulation) that *"for example for tax evasion, it is possible that when the source codes are publicized, that weak points will be discovered and exploited."* Participant 8711 further commented on statement 13 (see *table 5*) that *"transparency can be an obstacle for an investigation or for politically sensitive situations, by for example the media in public relations."* Participant 7982 also commented on statement 13, saying that *"[transparency] can lead to a stagnation of the operations of intelligence services."* Participant 9279 rejected statement 38 (on that citizens must have full control over their personal data), saying *"no, why should they? The government must have certain data under their control."* And, participant 7982 could align with statement 37 (on auditability for authorities but not civilians) because *"civilians do not always have enough understanding, and will act in their individual interests."*

The trade-off between transparency and performance was also considered in comments. For example, participant 9279 commented on statement 54 (on that the choice for an algorithm depends on its interpretability and explainability), saying *"no, the choice depends on how well it works and the integrity that it has."* A similar comment was echoed by participant 9279 on statement 16 (on transparency over effectiveness in a democratic society) that he *"[disagreed], because for a good use of AI, effectiveness is enormously important."* Participant 4280 also disagreed with statement 16, stating that this argument is *"a fallacy, lawyers language."* However, not all persons with this discourse pick performance over transparency. For example, participant 5636 commented on statement 33 (on explainability which limits algorithms) and 20 (on AI performance over transparency) that: *"this may not necessarily be a reason to not be transparent. It can sometimes help to understand that there is a correct handling of AI."* Participant 7982 further made a general comment at the end of the Q-sort that *"the attention for innovation is very important to me from the perspective of efficiency and improvement of quality, but this cannot happen at all costs. Control and monitoring by experts is necessary."*

An understanding for the need for transparency in some cases did come to the fore as well. For example, for the case of traceability, participant 7982 commented on statement 22 that *"[a robot] actually must be traceable, to make it possible that changes can be made."* Participant 7982 further commented on statement 06 (on algorithm traceability) that *"the general importance and the possibility to trace activities, and where needed to make adjustments, is very important to me."* Along similar lines, participant 9279 commented on statement 06 (on storing and tracing data back to a problem) and statement 09 (on transparency and accountability) that *"mistakes must be traceable to its origin. The responsibility must stay with humans."* Participant 8711 commented on statement 28 (on not allowing algorithms to be a black-box) that *"both companies and the government work with very sensitive intelligence. Civilians have the right to know what happens [with this intel], up until a certain degree."* Participant 8711 commented on statement 29 (see *table 5*) that:

> *openness of all results is important not only to evaluate but also for functionality. Otherwise a situation will arise like when a child only mentions his school passing grades. A developer/researcher will easily (certainly in the private sector) get pressure from above to produce positive research. Negative results will be easily held behind to maintain product trust and for stock values.*

4.6 Discourse C: reap the benefits of AI!

4.6.1 Discourse C interpretation based on statements with high salience (see table 6)
This discourse sees transparency as a blockade to AI development. They agree that when AI systems are not transparent that this facilitates the possibility to operate better than humans (20). As long as it does not impair AI developments, they agree that some elements of AI systems can be made transparent, such as storing data to facilitate tracing back where something went wrong (06), and the ability to control how personal data gets used (38). They find that openness would be too technical, that this would confuse people rather than generating more trust (08).

Persons with this discourse understand how transparency works. They understand that AI can reach its full potential without explainability (19). Therefore, they feel comfortable to trust an AI system even if it does not give a better explanation than a human (41). To them transparency refers to the 'what' of an AI system, this could include insight in algorithms or insight in data use (52). They do not let their go-to choice for an AI systems be guided by specific AI transparency factors, such as the need for interpretability or explainability (54).

Table 6: Discourse C high salience statements. Statements with * are discourse distinguishing statements at $p < 0.01$, with ** at $p < 0.05$.

Z-score	+5 (most like how I think)	Z-score	-5 (least like how I think)
2.356 **	20. The way that AI systems are self-aware and achieve goals is often not	-1.793 **	19. Without explainability of AI we cannot reach the potential of AI.

	transparent, and that is exactly why they are sometimes better than humans.		
1.766	06. Data could be stored, with which experts, like developers of the algorithm can trace back to why something went wrong.	-1.744	52. Transparency does not imply necessarily the insight in algorithms and data use.
1.725 **	38. Residents must maintain full control over their personal data.	-1.682 *	54. The choice of the algorithm depends on how well the algorithm can be interpreted and explained.
1.656 **	08. Openness about the algorithm is difficult because users often do not understand how such an algorithm works. Such an explanation is too complicated and does not contribute to trust.	-1.441	41. If the machine cannot give a better explanation than humans, then do not trust the machine.

4.6.2 Discourse C participant comments

Participants with this discourse often commented on performance and transparency. For example, participant 3832 commented on statement 20 (see *table 6*) that *"often patterns that can be seen / calculated by a computer are not obvious to the human way of thinking and can therefore only found by computers. Also, the computer is not "blinded" by emotions."* Participant 1592 also commented on statement 20 that *"it is often an incomprehensible black-box, but still better than a human."* Participant 1402 also commented on statement 20, saying that *"to model complex nonlinear and dynamic systems, we will arrive at models that are no longer fully transparent, but that does make correct forecasts. A machine makes predictions for which it uses information and modes of reasoning that are no longer interpretable by humans."* And to statement 19 (see *table 6*) participant 3832 commented that *"for whatever reason, it works. This happens so often in science and technology. If we had stopped for this reason then we would not be where we are today. Why does it work? Let's find that out later."* Participant 1592 also commented to statement 19, saying that *"AI does not need to be explainable to enjoy the positive results."* Participant 1592 further made the general remark that *"transparency is important but it cannot go too much against the progress of applicability, because the average human cannot understand it anyway."* Participant 1402 commented on statement 10 (on not allowing algorithms to be a black-box) that he *"thinks that you will always have a trade-off between complexity and transparency for AI, because the more complex something is, the less transparent it is.*

Participants also commented that transparency does not contribute to trust in the case of AI. For example, participant 3832 addressed statement 8 (see *table 6*) that *"you don't know how your car works exactly… but you can use it in the way it is supposed to be used. [The] same [applies] here."* And to statement 41 (see *table 6*) participant 3832 replied that *"it is often the case that it is better, the way the machine [decided]. But how should a machine explain that? Maybe the human did not give [the machine] the words to explain."* Participant 1592 replied to statement 18 (that black-box algorithms cause distrust) that

"it is indeed a black-box but not a source of distrust. People know almost nothing about the current technology that they use, like a microwave, but still they trust the result."

Participants are more generally in agreement with data transparency, to prevent harming privacy. For example, participant 3832 replied to statement 38 (see *table 6*) that *"it is way too often the case that people do not know what happens with their data. It should be a lot easier to look that up for every application."* Participant 1592 also commented on statement 38 that *"power cannot shift too far away from civilians."*

This discourse also agrees to make AI traceable. For example, participant 1592 replied to statement 06 (see *table 6*) that this *"ensures the improvement of algorithms, which is good."* In similar lines, participant 1592 commented on statement 23 (on logging AI decisions for auditing) that *"logged information can be useful for the next algorithm and it ensures for auditability to maintain a power balance."* Participant 6501 also commented on statement 06 saying that *"you should always be able to see where something goes wrong."*

Participants often did not find explainability a must. For example, participant 3832 commented on statement 54 (see *table 6*) that it is *"difficult. I might agree with interpretation. But only because you cannot explain the result of the algorithm does not mean that it is not correct."* Participant 1592 also commented to statement 54 that we *"must not use worse algorithms because we want to understand [algorithms]. A correct outcome weighs heavily."*

Participants often commented that it is not problematic if AI is not transparent, because people do not understand it or to prevent manipulation. Participant 3832 made the general remark that *"only because we do not fully understand how machines work, does not mean that we cannot use them. Look at science: we still don't understand a lot of what's going on in biology, chemistry and physics, and yet we are using it to make life better and easier. Why not do the same with AI? And find out how they work later.* Participant 1402 commented on statement 24 (that transparency leads to manipulation) that *"AI makes it possible to automate weapons, to circulate at large scales false information, to negatively influence systems. The threshold to do this is low, especially compared to nuclear weapons. A killer drone that automatically targets humans and shoots is in theory much easier to device."* Participant 1402 further commented on statement 10 (on manipulation) that *"if you understand how people respond to things, and how to keep them in your sphere of influence, then you can manipulate them, without them knowing about it. Think about social media and the endless recommendations people receive for example."*

Chapter 5: Analysis

In this chapter the, an in depth analysis is performed to unravel the meaning of AI transparency for each of the discourses. As mentioned in *chapter 2,* the meaning of the phenomena that is studied (AI transparency in this case) is a discourse. This meaning can be discovered by finding how persons ascribe ideas and classify the related AI transparency concepts. The theoretical underpinnings from *chapter 2* will be used to identify the meaning of AI transparency for each discourse. In addition, the empirical discourses are also compared with the hypothetical discourses from *chapter 2*. This would aid to identify where there is alignment with the current knowledge in the literature, and to identify new insights. It was expected that three discourses would surface. These expectations were confirmed in the factor analysis in *chapter 4*.

5.1 Discourse A: no excuses, we demand AI transparency!

Discourse A was found to be most compatible with the hypothesized "proponents" discourse. The application of the findings from *chapter 2* on the empirical findings revealed a deeper meaning of AI transparency for this discourse. This discourse often seeks for a high degree of transparency because they do not trust AI. This discourse considers transparency as a means to open the black-box, to generate understanding, accountability, and legitimacy. These factors are each argued to be contributors of trust. The trus argument can possibly explain the demand of this discourse for human-centric AI. Although this must be confirmed through further academic investigation. The new insights of this discourse are explained in the analysis below.

Legitimacy

Legitimacy is a topic of high salience for discourse A, and a topic with strong linkages to the "proponents" discourse. Statement 12 on legitimacy received the highest positive Z-score. Statement 27 with a factor array of +4 was also of salience. Transparency was linked to legitimacy because it would generate understanding. The concept of democracy was also added in this linkage. The transparency-legitimacy linkage was also said to contribute to trust.

These findings strongly match with the expectations for the proponents discourse. First, because, as mentioned in Meijer (2009), it was expected that *the added value to democracy* would be emphasized. The reference to *understanding* matters as well because this resonates with two benefits of transparency that are discussed by Weller (2017): understandability of decisions, and understanding of strengths and limitations. Understanding also matters because it could lessen ones "fear of the unknown" (Weller, 2017). Another benefit that was mentioned frequently by the participants, and mentioned by Weller (2017) was that transparency would generate trust. The linkage to accountability will be further discussed below.

Accountability

Accountability and the centrality of humans was another topic of high salience. For example, statement 9 (on transparency and that humans are accountable, and not the algorithm) was found to be more important (+5) than statement 53 (on transparency and that humans or the machine are accountable) (+3). It was reported that AI/algorithms must be traced to humans, that humans must be accountable, that humans must be in control, that humans must understand the algorithm. There were further remarks regarding a possible power imbalance and the existence of a dangerous situation. The transparency-legitimacy linkage

was also tied to accountability. As in that AI makers are responsible until AI is declared to be legitimately self-aware.

The first linkage that is found here is that accountability was expected to be an important benefit for the transparency proponents (de Laat, 2018; Meijer, 2009; Weller, 2017). The emphasis on traceability also strongly resonates with one of the benefits mentioned in Weller (2017).

What was less expected in comparison with the proponents discourse, was the emphasis on the centrality of humans vis-à-vis the AI system. The notion of human control here could possibly be partially explained by some benefits highlighted by Weller (2017): such as the ability to check and challenge the system, and his notion on transparency and verification. Control can also be linked with the "fear of the unknown" argument. For example, Buiten (2019, p. 3) mentions that "opacity of [AI] reinforces concerns about the uncontrollability of new technologies: we fear what we do not know." One argument could be that this discourse fears what would happen if humans are not in control. This ties for example to the fear for superintelligence as mentioned in Burton et al. (2017, p. 25) and Makridakis (2017, p. 51) which spell that losing control could mean an "existential threat" or that computers could "eventually [be] in charge of making all important decisions." The fear of superiority resonates with a comment of participant 8108 on statement 30 that *"it must not become the case that a robot becomes smarter than its maker."* And another comment of that sheds light on fear comes from participant 7722 that we could *"find ourselves in a dangerous situation."* Human-centric AI, human agency, and human oversight are themes which receive high attention in the scholarly field of AI (for example, see AI HLEG, 2019). But this is not specifically in the context of transparency. The provocation of arguments on human centrality indicates that there is an opportunity for further academic exploration on the relationship between transparency and human centrality.

Trade-offs

Another topic of high salience was on the trade-off between performance and transparency. The position of this discourse on the transparency-performance trade-off is at odds with discourse C. This discourse was found to choose transparency over performance. Whereas discourse C was found to choose performance over transparency.

Statement 20 (on that AI is better than humans because they are not transparent) was the lowest ranked statement of this discourse. One participant acknowledged this statement, and rejected it, saying that *"this does not mean that we must not try to understand it."* The transparency over performance position was bolstered by statement 33, the second lowest ranked statement (that explainability limits the added value of algorithms because it places on the complexity of algorithms). Participant 8108 replied to statement 33 that "*for people who do not have a high understanding about this subject, like myself, it must remain understandable."* Statement 16 (on that in a democratic society, effectiveness may not drop the obligation for transparency) was the fourth highest ranked statement of the discourse. Participant 6277 replied to statement 16 that this would be like *"being punished because something is less effective,"* and *"that in a democratic society it cannot be a reason to place efficiency in front of transparency."*

These findings can be linked back to the proponents discourse. The comment by participant 8108 ties back again to the notion of understanding, important benefits as outlined in Weller (2017, p.56). The

author specifically writes that it can help to "understand why one particular prediction or decision was reached" and to "understand how [the] system is working." The concept of democracy, brought to the fore by participant 6277, is as mentioned above another factor of importance for the proponents in Meijer (2009). The proponents' beliefs that transparency could bring more affluence to democracy (Meijer, 2009) are in this case threatened because, as one participant put it, this trade-off (statement 16) feels like *"being punished because something is less effective."*

Trust: the all encompassing factor?

Trust was another topic of high salience for this discourse. This was especially noticed from the high volume of comments on trust. Numerous comments were given on statements that specifically discussed trust. But trust was also in comments on statements that were not specifically about trust. Trust was often linked to understanding. One comment mentioned that everything should be as simple as possible, and another that explanations are needed for understanding and trust. The interconnectivity between trust and other points of salience for this discourse motivated a lengthier analysis here.

As discussed earlier in this section, trust and understanding were anticipated to be viewed as a transparency benefit by the proponents discourse. As mentioned in *chapter 2*, trust is one of the major obstacles to the development of AI (Rossi, 2019; Siau & Wang, 2018). Siau and Wang (2018, p. 50) make the explicit connection to transparency, understanding and trust; they write that "to trust AI applications, we need to understand how they are programmed and what function will be performed in certain conditions. This transparency is important, and AI should be able to explain/justify its behaviors." This explanation first allows to make the linkage that transparency is required to build understanding. And further that this understanding can be a precondition for trust. But a second relationship worth noting here is regards explaining and justifying behavior to earn trust. As mentioned in chapter 2, legitimacy in this study is understood as the "justifiability of a power relationship" (Eshuis & Edwards, 2013, p. 1071). This would then translate to explainability being a form to justify (to gain legitimacy) and that this in turn is a precondition for trust. This is in line with a comment by participant 5993 that *"you must be able to justify [the use of algorithms]... if you provide an explanation then you will… earn trust."*

To sum up, two findings here are important to note for this discourse. The first is that transparency facilitates understanding, which enhances trust. And the second is that explainability facilitates legitimacy (justification), which enhances trust. From this linkage, the unwillingness to accept black-boxes (as seen from the rejection of statement 32 on black-boxes) can also be clarified. As participant 6277 expressed: *"we mustn't want to get used to this… we must try to understand it's decisions."* The reference of the participant to understanding ties back to the relationship above, that making the black-box transparent could provide understanding. Which in turn could provide trust.

And finally, the salience that this discourse places on accountability could also be explained from the perspective of making black-boxes transparent and trust. As Rossi (2019, p. 128) writes: "issues include the black-box nature of some AI approaches… and the accountability and responsibility when an AI system is involved... Without answers to these [issues], many will not trust AI."

Based on these findings, the discourse could be said to be driven by trust-seeking behavior. Transparency facilitates the opening of the black-box, which can generate legitimacy, accountability, and

understanding. In turn, these factors combined promote trust. This might also explain the strong demand for human centrality. Because this discourse does not trust AI, they would not want to lose human control.

The argument of distrust could also explain the high volume of rejections to statements which favored opacity. The rejections were oftentimes loaded with suspicion. One participant (5993) said that *"a limited form of transparency for me means that you are not honest, and that you do not want to tell the truth."* Another participant said that *"we have no choice, we are neatly walking behind the elites."* And another participant (7722) mentioned that *"if it isn't traceable, then we will get a power imbalance, we must not want that."* The suspicious comments could explain why AI opacity generates distrust for this discourse, because: it generates a power imbalance, it keeps people inferior to the elites, and it is a sign of dishonesty.

Two participant remarks were all encompassing for the findings of this analysis. Participant 5993 said: *"I have the opinion that a shortage of knowledge / not knowing what something is / and not knowing what happens leads to distrust. The explanation of something to a human can lead to trust, it can give back trust."* And participant 3722 who remarked that: *"without transparency there cannot be trust."* And whether there is a glimpse of openness to opacity? In a convoluted way, the answer could be yes*: "algorithmic transparency can lead to a satisfactory level of trust which makes it possible that the right for secrecy can be kept."*

5.2 Discourse B: balance the needs for AI transparency!

Discourse B was found to be most compatible with the hypothesized "context-dependent" discourse. The application of the findings from *chapter 2* on the empirical findings revealed a deeper meaning of AI transparency for this discourse. This discourse often considers whether AI should or should not be transparent based on a utilitarian point of view. They consider the greater societal good to internally justify whether there is a need for transparency. This new insight is explained in the analysis below.

Context-dependence

A topic of high salience for discourse B was on the need to consider the context of a situation. High salience was given to statement 15 (that in some cases more limited forms of transparency could be considered) and to statement 21 (that there is a need to balance between transparency and commercial interests of secrecy).

This finding was strongly expected in the hypothesized "context-dependent" discourse. Hence the name of the hypothesized discourse "context-dependent." The argument for the need to consider the context of a case can be found in Buiten (2019). Regarding the statement 21, Buiten (2019, p. 18) writes that those who consider the context should "consider that the opacity of algorithms may have non-technical justifications, such as protecting trade secrets… [and] be aware of the costs of requiring transparency particularly for small companies, which may be disproportionately burdened." This remark sheds insight in some of the comments of the participants. For example, participant 5536 commented to statement 21 that *"a balance for the entire topic [of AI] is highly important. There is a large grey area. I understand the interests of both sides. You do not want to inflict economic damage."* And participant 9279

commented that *"companies and governments must also be able to do their work well to be of added value. A balance is therefore necessary."*

Utility

Buiten (2019, p. 18) further wrote on context-dependency that "if the costs of providing transparency are high, requiring it can potentially have negative effects on innovation." This remark resonates strongly with a comment by participant 7982 that *"the attention for innovation is very important to me from the perspective of efficiency and improvement of quality, but this cannot happen at all costs. Control and monitoring by experts is necessary."* What this comment illustrates is that the situation is carefully considered in terms of utility: innovation is important to the participant, but to a certain threshold (not at all costs). The red line would be crossed if innovation would mean having to give up control and monitoring by experts. An explanation for this mode of thinking is also reflected in Buiten (2019, p. 17) who writes that for context-dependency there is a "[need] to ask when it is worth requiring transparency, balancing the utility of transparency against the costs of generating such a system."

The utility argument can also be analyzed from the perspective of Zarsky (2016, p. 122), who writes that "even if transparency somewhat improved the accuracy of algorithmic processes, the aggregated costs of facilitating disclosure… render it costly. Once we acknowledge such factors, transparency does not appear to substantially enhance social welfare." Zarsky (2016) here closely weighs what transparency would mean in terms of enhancing social welfare. This is closely in line with the utilitarian way of thinking that "one ought to maximize the overall good - that is, consider the good of others as well as one's own good" (Driver, 2014).

Limited transparency if it's better for the greater good

The utilitarian mode of thinking can also be found in the arguments for limited transparency. For example, participant 5636 replied to statement 37 (that judges and authorities can audit, but civilians cannot) that "*there are serious parties needed who receive insight. Giving the civilians insight will only bring misery."* Participant 7982 also commented on statement 37 that *"civilians do not always have enough understanding, and will act in their individual interests."* From the utilitarian point of view, transparency is not of added value to the greater good if it makes persons act in their own interests, and neither is that the case if giving insights to civilians will "only bring misery."

The call for limited transparency is also reflected in the most salient statement of this discourse, statement 13 (that transparency would impair the effectiveness of intelligence operations). Participant 5636 for example agreed with this statement and replied that *"intelligence services must maintain the security of the society, this is important."* Participant 8711 replied to the same statement that *"transparency can be an obstacle for an investigation."* And participant 7982 replied that transparency *"can lead to a stagnation of the operations of intelligence services."* An explanation from a utilitarian point of view is that limited insights in the case of society is good because it ensures safety for society. Buiten (2019) also comments on this, with a reference to Wachter et al. (2017) that "regulatory transparency requirements should be context-dependent and based on risks to safety, fairness, and privacy."

Fitting to the overall findings for this discourse, two comments came to the fore. From participant 7982, who said that *"the common good, and the possibility to trace, and where needed to make adjustments, is*

what I find highly important." And from participant 8711, who said that *"transparency is necessary, but the level of [transparency] as well as insight for the masses, will always be dependent on the situation."*

5.3 Discourse C: reap the benefits of AI!

Discourse C was found to be most compatible with the hypothesized "opponents" discourse. The application of the findings from *chapter 2* on the empirical findings revealed a deeper meaning of AI transparency for this discourse. The biggest finding is that, contrary to what is written in the literature, that this discourse is often open to transparency if transparency does not affect the algorithm. This discourse is in favor of having algorithms whose performance is as robust as possible. Placing requirements on the transparency of the algorithm would affect the performance, and therefore they can be expected to be against this. Other forms of transparency that do not affect the performance of the algorithm could be accepted. These new insights are explained in the analysis below.

Opacity and performance

The most salient finding is that this discourse accepts opacity. This is in opposition to the position of discourse A, who rejects arguments that argue positively for opacity. The primary concern for discourse C is that the performance and development of AI systems are not affected by transparency. For example, participant 1592 replied to statement 54 (whether the choice of the algorithm depends on interpretability and explainability) that *"we must not use worse algorithms because we want to understand it. A correct outcome weighs heavily."* Participant 3832 further commented to statement 20 that *"often patterns that can be seen / calculated by a computer are not obvious to the human way of thinking and can therefore only found by computers. Also the computer is not "blinded" by emotions."* This verifies the expectation from the hypothesis that performance would be preferred over transparency. A good argument for this viewpoint can be found in Buiten (2019) that "there may be a trade-off between the capacity of a model and its explainability… By requiring more transparency, we may need to accept that the systems are less accurate than they could technically be… both clarity and accuracy are useful for preventing errors."

Transparency and maintained performance

One finding that came to the fore is contrary to what was expected in the hypothesis. With a second highest ranking for statement 6 (+5), discourse C is clearly open to some forms of transparency. Statement is regarding traceability, discourse C is in strong agreeance with discourse B that traceability is important. For example, participant 6501 commented on statement 6 (to store data to trace back to a mistake) that *"you must always be able to see where something goes wrong."* Participant 1592 further added on auditability that *"logged information can be useful for the next algorithm, it ensures auditability which can keep the power balance in check."* At first sight this acceptance towards transparency appears to be at odds with the former acceptance of opacity.

One possible explanation for this finding could be that discourse C does not see auditing or tracing as a threat to their needs to ensure robustness of the AI system. This could be in line with Lepri et al. (2018) who writes that auditing techniques can be used which examine the inputs and outputs of a system, but prevent that the system itself is revealed (Lepri et al., 2018). Buiten (2019) also reports on this, writing that the "tracing back how certain factors were used to reach an outcome" can be done by focusing on the input, or by focusing on the output of a system, something that "may be more feasible for programmers

than the approach focusing on the decision-model of the algorithm." Then what is the catch with the algorithm?

According to Strauß (2018, p. 4) AI algorithms are complex, and this can be explained because they "collect, analyse, de- and re- contextualise large data sets to explore and recognise patterns." The level of sophistication and complexity of an algorithm dictates the performance of an AI system (Strauß, 2018). Certainly, AI algorithms (as discussed in *chapter 2*) require computing power and data to function, but this does not affect the inherent performance capabilities of the algorithm at face value. From this perspective by only making the input and output of the AI system transparent (and not the algorithm) it could be prevented that the performance of the AI algorithm is affected. This aligns with the strong focus of the participants on well-functioning algorithms. Take for example the comment from participant 1592 on statement 6: *"this causes an improvement of the algorithms, that is good."* Note here that statement 6 is about "storing data" which means that it is input focused, and that the algorithm would stay intact. Another comment that can be exemplified is on statement 20 (on that AI systems are better than humans because they are not transparent) by participant 1592: *"it often is a black-box that is not understandable, but it is still better than humans."* Based on the argument above, it could be interpreted that the discourse understands that the AI algorithm is indeed a black-box (because it is complex), and that it is this complexity that makes it better than humans. One comment by participant 1402 on statement 20 reflects this line of thought:

> *to model highly complex, non-linear, dynamic systems, we will make systems that are no longer transparent but that do make the right predictions. The machine will make predictions based on processes and reasons that are no longer interpretable by humans.*

Transparency and distrust

Because this discourse understands that AI algorithms are complex, their fourth highest ranking (+5) of statement 08 (on that openness about the algorithm is difficult because users often do not understand how such an algorithm works. Such an explanation is too complicated and does not contribute to trust) can also be better understood. For example, participant 3832 commented on statement 8 that *"you don't know how your car works exactly… but you can use it in the way it is supposed to be used. Same here."* The disagreement with statement 18 (on that an algorithm is for many a mysterious black-box and that causes distrust) by participant 1582 was in a similar vein: *"it is indeed a black-box, but not a source of distrust. People know almost nothing of the current technology that they use, like a microwave, but still they trust the result."* From Weller's (2017) point of view, the complexity of the system ultimately could be something that people do not want to hear, it could be a "harsh truth" which could affect trust. And from the perspective of Buiten (2019, p. 18) another explanation could be given:

> Regardless, for any transparency requirements we need to consider whether a technically feasible explanation would be useful for the prospective recipients. Merely bombarding people with information they cannot process will not help them to make better decisions. Neither will a requirement for producers of algorithms to hand over the code help courts in assigning liability."

The findings of this analysis illustrates that this discourse considers whether transparency affects the algorithm. They are concerned that transparency places demands on the algorithm that will undermine the

performance of the AI system. They can agree to transparency regarding non-algorithm components such as data-use, because this is not perceived to affect performance. This discourse understands the inherent complex nature of AI algorithms, they therefore fear that transparency of the algorithms would cause distrust rather than trust. A fitting closing statement for this discourse is taken from participant 1592:

> *Transparency is important but it cannot go against the applicability [of AI], because the average person cannot understand it anyways. In specific cases it is allowed that a higher value is given to [transparency] when the safety of the user is at stake (government and sensitive/private data).*

Conclusion

The arrival of AI applications across many domains of ordinary life has made one thing clear: AI is here to stay. Some applications have even surpassed Turing's (1950) once contemplated possibilities on learning and thinking machines. The growing robustness of AI is now raising questions whether artificial intelligence (AI) will, at some point surpass human intelligence. The entry of AI in society has caused many to consider the risks and benefits of its applications. This has sparked ethical debates about what AI should and should not be allowed to do, and what role it is supposed to have for society. The ethical debates in a seemingly unregulated field have fueled diverse stakeholder groups to call for policymakers to act. Ursula von der Leyen, President-elect of the European Commission, has expressed that she would respond to the need for legislation in her first 100 days of office. In the Netherlands, regulating AI has become a top priority in the Digital Strategy of the Dutch government, with a variety of policy responses in the pipeline.

One topic that is of salient debate is regarding the 'black-box' or 'opaque' nature of AI. The AI transparency debates are heated because there appears to be disagreement over if, to what extent, and to whom AI applications should be transparent. This study anticipates that disagreement over AI transparency can pose a barrier to policymakers in the Netherlands and the EU to devise policy that matches with the concerns of the public. To identify opportunities to move forward, this study aims to unravel the discourses on AI transparency. On this backdrop, and with the case of the study being the Netherlands, the following research question was raised: what are the public discourses of AI transparency in the Netherlands?

To answer this research question, the first part of this study reviewed the existing literature on transparency, AI, and AI transparency more specifically. The theoretical and empirical observations from the literature resulted in a conceptual framework surrounding AI transparency. Five main enabling factors of AI transparency were identified: explainability, interpretability, auditability, traceability, and communication. These factors influenced the levels of AI transparency, which were classified into three levels of AI transparency: full transparency, limited transparency, and black-box/opacity. The different levels of AI transparency were found to eight main effects, on: accountability, legitimacy, trust, fairness, fear, privacy, trade-off effects, perverse effects. The conceptual framework and a review on discourses of AI transparency subsequently aided to hypothesize three AI transparency discourses: 1) the "proponents," 2) the "opponents," and 3) it is "context-dependent."

The subsequent part of the study set out to answer the research question based on empirical data. The Q-methodology was used to empirically measure the discourses from a sample of 31 residents in the Netherlands. The participants were asked to 'Q-sort' 54 statements on a quasi-normal distribution grid, spanning from "most how I think" to "least how I think." To ensure that the diversity of the AI transparency topic was represented, the selection of statements was based on predetermined selection criteria, and based on the conceptual framework. Factor analysis using principal component analysis and varimax factor rotation was used for the statistical analysis of the 31 Q-sorts. Three discourses produced the most robust statistical result for the 31 Q-sorts. The results were subsequently compared with the hypothesized discourses, to identify alignment with the literature and to review new insights. And, the results were analyzed to discover the deeper discourse meaning using theoretical findings from *chapter 2*.

In response to the research question, based on the research findings and the analysis, the three public discourses of AI transparency found in the Netherlands in a sample of 31 participants are as follows:

Discourse A: no excuses, we demand AI transparency! This discourse was found to be strongly in favor of AI transparency. For them, transparency is a means to generate the benefits of understanding, legitimacy, accountability. In-depth analysis revealed that these benefits were of importance to trust AI. The in-depth analysis also revealed that AI must be human-centric, even if AI is transparent. To them, transparency can prevent power imbalances between societal groups and is an important feat in a democratic society. Arguments for AI opacity were refuted, and oftentimes with remarks of suspicion. This discourse was closely related to the hypothesized "proponents" discourse.

Discourse B: balance the needs for AI transparency! This discourse was found to consider the need for an AI transparency depending on the context of the situation. To them, it is often about understanding and balancing the needs of different stakeholder groups. In-depth analysis revealed that this discourses often considers cases from a utilitarian point of view. If, to what extent, and to whom AI applications should be transparent often depended on what was better for the greater good. This discourse was closely related to the hypothesized "context-dependent" discourse.

Discourse C: reap the benefits of AI! This discourse was found to favor AI opacity, but not as strongly as suggested in the literature. To them, transparency is frowned upon because they worry that it would affect the performance and development of AI systems. They understand how complex AI algorithms are, and that being transparent could spark distrust among the general public. This discourse was comparable to the hypothesized "opponents" discourse, but in-depth analysis revealed that there were distinct differences. The most notable difference is that this discourse is open to various forms of AI transparency. Transparency is generally accepted if it does not affect the performance of the AI system. This would mean that input or output of a system could be made transparent, as long as the algorithm (which determines the performance) is not tampered with.

Theory and added value

The theory of this study enshrined in the conceptual framework was robust enough to execute this study. First, during statement classification and selection, the framework proved to be capturing the breadth of the AI transparency topic. No additional sub-categories were found during statement selection that was missing in the framework. The only additional step that was taken was to internally split the "fear" and "fairness" sub-categories to represent the diversity of statements well enough. The theory underpinning the conceptual framework was also robust enough for the in-depth analysis of the empirical results. However, some distinct elements were missing. The first is that the concept of utilitarianism was not discussed or taken on board in the theory chapter. The second is that accountability, legitimacy, and understanding were found to be mediating variables that would generate trust for discourse A. In the framework, transparency and trust are directly linked. Moreover, discourse A revealed a reverse relationship between trust and transparency. For discourse A, distrust was the factor that could explain the demand for AI transparency. The third missing element is that the theory did not treat human-centric AI. Human-centric AI is a large field on its own in the AI literature, yet it was not expected to surface this strongly in one of the discourses. Further research is needed to investigate the transparency and human-

centricity link. Overall, it was found that discourse A wants AI to be human-centric, even when AI is transparent. A final missing element is that Discourse C matched most closely with the "opponents" discourse, but there were striking differences that were not accounted for in the "opponents" arguments in the literature. The existing conceptual framework, however, could still decipher their biggest concern. Which was that discourse C was worried that transparency would tamper with AI algorithms, which would affect AI performance. If this was not the case, discourse C was found to be open to transparency. The conceptualization of the AI concept proved very helpful here. It was already conceptualized that "complex algorithms" were a distinct attribute of the concept. This helped to dissect the line of thinking of discourse C.

The real added value to the theory is thus: the robustness of the conceptual framework; the suitability of the AI concept to explain what AI is; the insights regarding the mediating variables to trust for discourse A; the insights regarding the utilitarian view for discourse B; and the insights regarding the openness to transparency for discourse C (as long as AI performance and development is not affected).

Strengths and weaknesses

The first strength of this research is the robust statistical analysis criteria were applied, which helped to identify the presence of the three discourses that are discussed above. Further, the robust analysis criteria ensured that confounding variables were not included in the discourse findings. The second strength is that the sampling method in terms of demographic background did not introduce bias into the research. The diversity can be seen by the demographic composition for each discourse in appendix 7. A third strength is that this thesis fills a knowledge gap in the field of AI discourses, and AI transparency discourses. This study can set in motion subsequent academic inquiry on AI discourses. In addition, it fills a knowledge gap of Q-methodology studies on AI. This study could serve as a template for future Q-methodology studies on AI.

The weakness of this research is that snowball sampling may not have been the best sampling method that could have been chosen. Because snowball sampling is a non-probability technique, the participant selection was based on my judgment. This can introduce a sampling bias, which can become more significant with more participants. However, I took specific steps, especially regarding the demographics of participants, to prevent bias from seeping into the study. Another weakness of this study is that the Q-methodology captures the 'internal standpoint' of the respondents. It measures subjectivity. This means that this study cannot capture the 'external standpoint' of the researcher, which is often possible in a large-N survey study. A large-N survey study could be devised to validate the findings of this research externally. Nevertheless, the added value of the Q-methodology is that it can capture subjectivity and then quantitatively analyze this subjectivity. This cannot be achieved through a large-N survey. Therefore, the two study methods can be complementary to each other.

Research suggestions

The first research suggestion would be to conduct a large-N survey to validate the results of this study externally. This could make the results of this study generalizable to the general population in the Netherlands. The second research suggestion, as discussed above is to investigate the connection between AI transparency and human-centric AI further. This connection did not feature prominently in the theory part of this study, but it was an important feature for the most dominant discourse. The third research

suggestion is to investigate the openness of the "opponents" discourse to transparency further. Such an investigation could further map which types of AI transparency the discourse is open to, which can be useful for policymakers to devise AI transparency legislation that can be accepted by the most dominant discourses. The fourth research suggestion would be to further investigate the relationship between trust and AI transparency. For the most dominant discourse, discourse motivated the demand for transparency. Transparency in turn was necessary to generate understanding, accountability, and legitimacy. These factors in turn generated trust. Trust was a pressing issue for this discourse, to decipher how trust can be generated for the discourse can be valuable to all stakeholders that work with AI. Finally, the last research suggestion would be to look further into the utilitarian perspective of the context-dependent discourse. The utilitarian perspective is briefly discussed in the literature, but not prominently. A further in depth investigation on the utilitarian perspective could possibly chart the cognitive reasoning on AI transparency for the context-dependent discourse better.

Policy recommendations

This research could be helpful for policymakers to devise AI transparency policy that could be acceptable to different discourses of AI transparency. As identified, all discourses in this study are open to some form of transparency. This means that there is terrain to legally place some form of transparency requirements on AI. One of the main points of friction among the discourses regards making AI transparent in such a manner that it would not affect the performance of AI. Especially the discourse that is least in favor of transparency worries that transparency affects the performance of AI algorithms. This sentiment was also recorded among some of the persons who aligned with the context-based discourse. For the context-based discourse, their worries were especially regarding the cases where accuracy is highly important. These worries could be avoided by making AI transparent in a manner that it does not affect its performance. The context-based discourse moreover was found to often consider the need for AI based what was best for the greater good. They are open to transparency or opacity if it is the best option for society. Another finding to consider is that the most dominant discourse is strongly in favor of AI transparency. They are highly skeptical towards any arguments that attempt to nuance the need for transparency. They strongly reject arguments which speak positively of AI opacity. Moreover, for this discourse, transparency does not imply that AI systems can be held accountable, or that AI can be in charge. Instead, they find that even if AI is transparent, it must remain human-centric.

References

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

AI HLEG. (2019). *High Level Expert Group on AI: Ethics Guidelines for Trustworthy AI*. Brussels.

AINED. (2018). *AI for the Netherlands: Enlarge Accelerate and Connection*. Retrieved from https://order.perssupport.nl/file/pressrelease/a253ded8-1863-4d07-b79d-72639baefc51/6be5dcdc-b9d8-4675-bd80-9425826c840c/AIvNL20181102final.pdf

Anastasopoulos, L. J., & Whitford, A. B. (2018). *Machine Learning for Public Administration Research, with Application to Organizational Reputation*. *SSRN*. https://doi.org/10.2139/ssrn.3178287

Araujo, T., de Vreese, C., Helberger, N., Kruikemeier, S., van Weert, J., Bol, N., … Taylor, L. (2019). *Automated Decision-Making Fairness in an AI-driven World: Public Perceptions, Hopes and Concerns*. Amsterdam. Retrieved from https://www.ivir.nl/publicaties/download/Automated_Decision_Making_Fairness.pdf

Barry, J., & Proops, J. (1999). Seeking sustainability discourses with Q methodology. *Ecological Economics*, *28*(3), 337–345. https://doi.org/10.1016/S0921-8009(98)00053-6

Brkan, M. (2019). Do Algorithms Rule the World? Algorithmic Decision-Making in the Framework of the GDPR and Beyond. *International Journal of Law and Information Technology*, *0*, 1–31. https://doi.org/10.2139/ssrn.3124901

Brown, A. J., Vandekerckhove, W., & Dreyfus, S. (2014). The relationship between transparency, whistleblowing, and public trust. In *Research Handbook on Transparency* (pp. 30–58). Edward Elgar Publishing. https://doi.org/10.4337/9781781007945.00008

Brown, S. R. (1980). *Political Subjectivity: Application of Q Methodology in Political Science*. London: Yale University Press.

Brown, S. R. (1993). A primer on Q methodology. *Operant Subjectivity*, *16*(3/4), 91–138.

Brown, S. R., Durning, D. W., & Selden, S. C. (2008). Q Methodology. In G. J. Miller & K. Yang (Eds.), *Handbook of Research Methods in Public Administration* (2nd ed., pp. 721–755). New York: Auerbach Publications.

Buiten, M. C. (2019). Towards Intelligent Regulation of Artificial Intelligence. *European Journal of Risk Regulation*, *0*(0), 1–19. https://doi.org/10.1017/err.2019.8

Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., & Walsh, T. (2017). Ethical Considerations in Artificial Intelligence Courses. *AI Magazine, Association for the Advancement of Artificial Intelligence*, 22–34. Retrieved from http://arxiv.org/abs/1701.07769

Calo, R. (2017). Artificial Intelligence Policy: A Primer and Roadmap. *SSRN*, *51*, 399–435. https://doi.org/10.2139/ssrn.3015350

Casalicchio, G., Molnar, C., & Bischl, B. (2019). Visualizing the feature importance for black box models. In *Machine Learning and Knowledge Discovery in Databases: European Conference* (pp. 655–670). Dublin. https://doi.org/10.1007/978-3-030-10925-7_40

Casares, A. P. (2018). The brain of the future and the viability of democratic governance: The role of artificial intelligence, cognitive machines, and viable systems. *Futures*, *103*, 5–16. https://doi.org/10.1016/j.futures.2018.05.002

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. *Science and Engineering Ethics*, *24*, 505–528. https://doi.org/10.1007/s11948-017-9901-7

Coogan, J., & Herrington, N. (2011). Q methodology: an overview. *Research in Secondary Teacher Education*, *1*(2), 24–28.

Cormen, T. H. (2013). *Algorithms Unlocked*. London: The MIT Press.

de Cremer, D. (1999). Trust and Fear of Exploitation in a Public Goods Dilemma. *Current Psychology*, *18*(2), 153–163. https://doi.org/10.1007/s12144-999-1024-0

Cucciniello, M., Porumbescu, G. A., & Grimmelikhuijsen, S. (2017). 25 Years of Transparency Research: Evidence and Future Directions. *Public Administration Review*, *77*(1), 32–44. https://doi.org/10.1111/puar.12685

Desouza, K. C., & Jacob, B. (2017). Big Data in the Public Sector: Lessons for Practitioners and Scholars. *Administration and Society*, *49*(7), 1043–1064. https://doi.org/10.1177/0095399714555751

Driver, J. (2014). The History of Utilitarianism. *The Stanford Encyclopedia of Philosophy*, *Winter edi*. Retrieved from https://plato.stanford.edu/archives/win2014/entries/utilitarianism-history/

Dryzek, J. S., & Berejikian, J. (1993). Reconstructive democratic theory. *The American Political Science Review*, *87*(1), 48–60. Retrieved from https://search.proquest.com/docview/214433777?accountid=10267

Eghbalighazijahani, A., Hine, J., & Kashyap, A. (2013). How To Do A Better Q-Methodological Research : A Neural Network Method For More Targeted Decision Making About The Factors Influencing Q-Study. In *ITRN 2013, 5-6th September, Trinity College Dublin* (pp. 1–11).

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. (2019). Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. In *International Conference on Intelligent User Interfaces* (pp. 1–13). Retrieved from http://arxiv.org/abs/1901.03729

Eshuis, J., & Edwards, A. (2013). Branding the City: The Democratic Legitimacy of a New Mode of Governance. *Urban Studies*, *50*(5), 1066–1082. https://doi.org/10.1177/0042098012459581

Etikan, I., & Bala, K. (2017). Sampling and Sampling Methods. *Biometrics & Biostatistics International Journal* , *5*(6), 149–151. https://doi.org/10.15406/bbij.2017.05.00149

EU Parliament, & Council of the EU. The protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), Pub. L. No. Regulation (EU) 2016/679, 1 (2016). EUR-Lex. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1528874672298&uri=CELEX%3A32016R0679

European Commission. (n.d.). My rights on data protection. Retrieved August 8, 2019, from https://ec.europa.eu/info/law/law-topic/data-protection/reform/rights-citizens/my-rights_en

European Commission. (2018). EU Member States sign up to cooperate on Artificial Intelligence. Retrieved August 6, 2019, from https://ec.europa.eu/digital-single-market/en/news/eu-member-states-sign-cooperate-artificial-intelligence

European Council. (2019a). *A new strategic agenda 2019-2024*. Brussels. Retrieved from https://www.consilium.europa.eu/media/39914/a-new-strategic-agenda-2019-2024.pdf

European Council. (2019b). *Leaders' Agenda: Strategic Agenda 2019-2024 Outline*. Brussels. Retrieved from https://www.consilium.europa.eu/media/39291/en_leaders-agenda-note-on-strategic-agenda-2019-2024-0519.pdf

van Exel, J., & de Graaf, G. (2005). *Q methodology: A sneak preview*. Retrieved from http://reserves.library.kent.edu/courseindex.asp;

Fairbanks, J., Plowman, K. D., & Rawlins, B. L. (2007). Book Review The Handbook of Public Affairs. *Journal of Public Affairs*, *7*, 23–37. https://doi.org/10.1002/pa

Filgueiras, F. (2016). Transparency and accountability: principles and rules for the construction of publicity. *Journal of Public Affairs*, *16*(2), 192–202. https://doi.org/10.1002/pa.1575

Frantzi, S., Carter, N. T., & Lovett, J. C. (2009). Exploring discourses on international environmental regime effectiveness with Q methodology: A case study of the Mediterranean Action Plan. *Journal of Environmental Management*, *90*, 177–186. https://doi.org/10.1016/j.jenvman.2007.08.013

Gasper, D., & Apthorpe, R. (1996). Introduction: Discourse analysis and policy discourse. *The European Journal of Development Research*, (special issue), 1–15. https://doi.org/10.1080/09578819608426650

Gasser, U., & Almeida, V. A. F. (2017). A Layered Model for AI Governance. *IEEE Internet Computing*, *21*(6), 58–62. https://doi.org/10.1109/MIC.2017.4180835

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." *AI Magazine, Association for the Advancement of Artificial Intelligence*, 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gössling, S., Cohen, S. A., & Hares, A. (2016). Inside the black box: EU policy officers' perspectives on transport and climate change mitigation. *Journal of Transport Geography*, *57*, 83–93. https://doi.org/10.1016/j.jtrangeo.2016.10.002

Government Office for Science. (2015). *Artificial intelligence: opportunities and implications for the future of decision making*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf

Grimmelikhuijsen, S., Porumbescu, G., Hong, B., & Im, T. (2013). The Effect of Transparency on Trust in Government: A Cross-National Comparative Experiment. *Public Administration Review*, *73*(4), 575–586. https://doi.org/10.1111/puar.12047

Hajer, M., & Versteeg, W. (2005). A decade of discourse analysis of environmental politics: Achievements, challenges, perspectives. *Journal of Environmental Policy and Planning*, *7*(3), 175–184. https://doi.org/10.1080/15239080500339646

Hale, T. N. (2008). Transparency, Accountability, and Global Governance. *Global Governance*, 73–94. https://doi.org/10.1177/0268580903018002001

Hegelich, S. (2017). Deep learning and punctuated equilibrium theory. *Cognitive Systems Research*, *45*, 59–69. https://doi.org/10.1016/j.cogsys.2017.02.006

Helbing, D. (2018). Machine Intelligence: Blessing or Curse? It Depends on Us! In D. Helbing (Ed.), *Towards Digital Enlightenment : Essays on the Dark and Light Sides of the Digital Revolution* (pp. 25–39). Cham: Springer.

Hermelingmeier, V., & Nicholas, K. A. (2017). Identifying Five Different Perspectives on the Ecosystem Services Concept Using Q Methodology. *Ecological Economics*, *136*, 255–265. https://doi.org/10.1016/j.ecolecon.2017.01.006

Jan, B., Farman, H., Khan, M., Imran, M., Islam, I. U., Ahmad, A., … Jeon, G. (2019). Deep learning in big data Analytics: A comparative study. *Computers and Electrical Engineering*, *75*, 275–287. https://doi.org/10.1016/j.compeleceng.2017.12.009

Janssen, M., & van den Hoven, J. (2015). Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? *Government Information Quarterly*, *32*, 363–368. https://doi.org/10.1016/j.giq.2015.11.007

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*, 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI Discourse. *Minds and Machines*, *27*, 575–590. https://doi.org/10.1007/s11023-017-9417-6

Johnson, J., Denning, P., Delic, K. A., & Sousa-Rodrigues, D. (2018). Big Data or Big Brother? That is the question now. *Ubiquity*, (August), 1–10. https://doi.org/10.1145/3158352

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, *62*(1), 15–25. https://doi.org/10.1016/j.bushor.2018.08.004

Kenward, L. (2019). *Developing a framework as a concourse management strategy: a selected literature review to guide novice researchers using Q Methodology*. *Nurse Researcher*. Cumbria. https://doi.org/10.7748/nr.2019.e1616

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., … Hadsell, R. (2017). Measuring Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

de Kleer, B. (2019). *Understanding "what" people think about European Integration Q-methodological Study of Dutch' Discourses on the Freedom of Movement of Persons*. Leiden University.

Kosack, S., & Fung, A. (2014). Does Transparency Improve Governance? *Annual Review of Political Science*, *17*, 65–87. https://doi.org/10.1146/annurev-polisci-032210-144356

de Laat, P. B. (2018). Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy and Technology*, *31*, 525–541. https://doi.org/10.1007/s13347-017-0293-z

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy and Technology*, *31*, 611–627. https://doi.org/10.1007/s13347-017-0279-x

von der Leyen, U. (2019). *A Union that strives for more: My agenda for Europe*. Retrieved from https://ec.europa.eu/commission/sites/beta-political/files/political-guidelines-next-commission_en.pdf

Lorscheid, I., Heine, B. O., & Meyer, M. (2012). Opening the "Black Box" of Simulations: Increased Transparency and Effective Communication Through the Systematic Design of Experiments. *Computational and Mathematical Organization Theory*, *18*, 22–62. https://doi.org/10.1007/s10588-011-9097-3

Mabillard, V., & Pasquier, M. (2016). Transparency and trust in government (2007–2014): A comparative study. *The NISPACEE Journal of Public Administration and Policy*, *9*(2), 69–92.

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

Meijer, A. (2009). Understanding modern transparency. *International Review of Administrative Sciences*, *75*(2), 255–269. https://doi.org/10.1177/0020852309104175

Meijer, A., Hillebrandt, M., Curtin, D., & Brandsma, G. J. (2010). Open Government: Connecting Discourses on Transparency and Participation. In *NIG Conference 2010 Good Governance Colloquium* (pp. 1–23). Maastrict.

Merriam-Webster. (2019a). Definition of Intelligence by Merriam-Webster. Retrieved August 7, 2019, from https://www.merriam-webster.com/dictionary/intelligence

Merriam-Webster. (2019b). Definition of Transparency by Merriam-Webster. Retrieved August 7, 2019, from https://www.merriam-webster.com/dictionary/transparency?utm_campaign=sd&utm_medium=serp&utm_source=jsonld

Merriam-Webster. (2019c). Definition of Transparent by Merriam-Webster. Retrieved August 7, 2019, from https://www.merriam-webster.com/dictionary/transparent

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, (July-December), 1–21. https://doi.org/10.1177/2053951716679679

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. In *Conference on Fairness, Accountability, and Transparency* (p. 10). Atlanta. GA. https://doi.org/10.1145/3287560.3287574

Moore, J. D., & Wiemer-Hastings, P. (n.d.). *Discourse in Computational Linguistics and Artificial Intelligence*. Edinburgh.

Moore, S. (2018). Towards a Sociology of Institutional Transparency: Openness, Deception and the Problem of Public Trust. *Sociology*, *52*(2), 416–430. https://doi.org/10.1177/0038038516686530

Nabi, R. L. (2003). Exploring the Framing Effects of Emotion. *Communication Research*, *30*(2), 224–247. https://doi.org/10.1177/0093650202250881

Newman, I., & Ramlo, S. (2015). Using Q Methodology and Q Factor: Analysis in Mixed Methods Research. In A. Tashakkori & C. Teddlie (Eds.), *SAGE Handbook of Mixed Methods in Social & Behavioral Research* (pp. 505–530). Thousand Oaks: SAGE Publications Inc. https://doi.org/10.4135/9781506335193.n20

Niedziałkowski, K., Komar, E., Pietrzyk-Kaszyńska, A., Olszańska, A., & Grodzińska-Jurczak, M. (2018). Discourses on Public Participation in Protected Areas Governance: Application of Q Methodology in Poland. *Ecological Economics*, *145*, 401–409. https://doi.org/10.1016/j.ecolecon.2017.11.018

Papadopoulos, Y. (2010). Accountability and Multi-level Governance: More Accountability, Less Democracy? *West European Politics*, *33*(5), 1030–1049. https://doi.org/10.1080/01402382.2010.486126

Pesapane, F., Volonté, C., Codari, M., & Sardanelli, F. (2018). Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights into Imaging*, *9*, 745–753. https://doi.org/10.1007/s13244-018-0645-y

Rajaraman, V. (2014). John McCarthy-Father of Artificial Intelligence. *Resonance*, (March), 198–207.

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges BT - Explainable and Interpretable Models in Computer Vision and Machine Learning. In H. J. Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, & M. A. J. van Gerven (Eds.), *Explainable and Interpretable Models in*

Computer Vision and Machine Learning (pp. XVII, 299). Springer International Publishing. https://doi.org/10.1007/978-3-319-98131-4_2

Rhoads, J. C. (2014). Q Methodology. In *SAGE Research Methods Cases* (pp. 1–18). London: SAFE Publications Ltd. https://doi.org/10.4135/978144627305014534166

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, *1*, 33–36. https://doi.org/10.1002/hbe2.117

Rijksoverheid. (2019a). *Dutch Digitization strategy 2.0: Here it is possible. It happens here*.

Rijksoverheid. (2019b). The Netherlands Digital: agreements for better cooperation digitization. Retrieved August 6, 2019, from https://www.rijksoverheid.nl/actueel/nieuws/2019/03/21/nederland-digitaal-afspraken-voor-betere-samenwerking-digitalisering

Risse, T., & Kleine, M. (2007). Assessing the legitimacy of the EU's treaty revision methods. *Journal of Common Market Studies*, *45*(1), 69–80. https://doi.org/10.1111/j.1468-5965.2007.00703.x

Rossi, F. W. (2019). Building Trust in Artificial Intelligence. *Journal of International Affairs*, *72*(1), 127–134. https://doi.org/10.2307/26588348

Rousseau, D. M. ., Sitkin, S. B. ., Burt, R. S. ., & Camerer, C. (1998). Introduction to Special Topic Forum : Not so Different after All : A Cross-Discipline View of Trust. *Academy of Management Review1*, *23*(3), 393–404.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). *Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models*. Retrieved from http://arxiv.org/abs/1708.08296

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *The 64th Annual Meeting of the International Communication Association* (pp. 1–23). Seattle, WA. Retrieved from https://pdfs.semanticscholar.org/b722/7cbd34766655dea10d0437ab10df3a127396.pdf

Schmidt, V. A. (2012). *Democracy and Legitimacy in the European Union*. (E. Jones, A. Menon, & S. Weatherill, Eds.), *Oxford Handbooks Online*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199546282.013.0046

Schmolck, P. (2014). The Q Method Software. Retrieved from http://schmolck.org/qmethod/index.htm#PQMethod

Sharma, G. (2017). Pros and cons of different sampling techniques. *International Journal of Applied Research*, *3*(7), 749–752. Retrieved from www.allresearchjournal.com

Shemmings, D., & Ellingsen, I. T. (2014). Using Q methodology in qualitative interviews. In J. F. Gubrium, J. A. Holstein, A. B. Marvasti, & K. D. McKinney (Eds.), *The SAGE Handbook of Interview Research: The Complexity of the Craft* (pp. 415–426). Thousand Oaks: SAGE Publications Inc. https://doi.org/10.4135/9781452218403.n29

Shreck, B., & Vedlitz, A. (2016). The Public and Its Climate: Exploring the Relationship Between Public Discourse and Opinion on Global Warming. *Society & Natural Resources*, *29*(5), 509–524. https://doi.org/10.1080/08941920.2015.1095380

Siau, K., & Wang, W. (2018). Building Trust in Artificial Intelligence,Machine Learning and Robotics. *Cutter Business Technology Journal*, *31*(2), 47–53. Retrieved from https://ezproxy.southern.edu/login?qurl=http%3A%2F%2Fsearch.ebscohost.com%2Flogin.aspx%3Fdirect%3Dtrue%26db%3Da9h%26AN%3D134748798%26site%3Dehost-live%26scope%3Dsite

Sluis, L. F. (2018). *Public Discourses in the Netherlands on the Freedom of Movement in the European Union: A Q-Methodological Approach*. Leiden University.

Smith, A. (2018). *Public attitudes toward computer algorithms*. Retrieved from http://www.pewinternet.org/2018/11/16/algorithms-in-action-the-content-people-see-on-social-media/

Stephenson, W. (1953). *The Study of Behavior: Q-technique and Its Methodology*. Chicago: The University of Chicago Press.

Stiglitz, J. E. (1999). On Liberty, the Right to Know, and Public Discourse: The Role of Transparency in Public Life. In *The Oxford Amnesty Lecture Series* (p. 155). Oxford.

Strauß, S. (2018). From Big Data to Deep Learning: A Leap Towards Strong AI or 'Intelligentia Obscura'? *Big Data and Cognitive Computing*, *2*(3), 16–35. https://doi.org/10.3390/bdcc2030016

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, *36*, 368–383. https://doi.org/10.1016/j.giq.2018.09.008

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science*, *361*(6404), 751–752. https://doi.org/10.1126/science.aat5991

Thomas, D. M., & Watson, R. T. (2002). Q-sorting and MIS Research: A Primer. *Communications of the Association for Information Systems*, *8*, 141–156. https://doi.org/10.17705/1cais.00809

TNS. (2017). *Attitudes towards the impact of digitisation and automation on daily life March 2017 May 2017*. Brussels.

Toshkov, D. (2016). *Research Design in Political Science*. Macmillan International Higher Education.

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*, 105–112. https://doi.org/10.1007/s10676-009-9187-9

Turing, A. M. (1950). Computing Machinery and Intelligence. *Computing Machinery and Intelligence*, *Mind 49*, 433–460. https://doi.org/10.1016/B978-0-12-386980-7.50023-X

Tweede Kamer. (2018a). *Motion of the member Verhoeven et al. About the public nature of the operation and the source code of algorithms and analysis methods*.

Tweede Kamer. (2018b). *Research into the use of algorithms within the government* (Vol. 0000967264). Den Haag.

Tweede Kamer. (2018c). *Transparency of algorithms used by the government*. Den Haag.

Veale, M., & Brass, I. (2019). Administration by Algorithm? Public Management meets Public Sector Machine Learning. In K. Yeung & M. Lodge (Eds.), *Algorithmic Regulation* (pp. 1–30). Oxford University Press. Retrieved from https://ssrn.com/abstract=3375391

Verhue, D., & Mol, P. (2018). *An investigation into the knowledge and attitude of citizens and entrepreneurs regarding artificial intelligence*. Den Haag.

VNG. (2018). Artificial intelligence opportunity for municipalities: Civil servant remains in charge | VNG. Retrieved August 6, 2019, from https://vng.nl/kunstmatige-intelligentie-kans-voor-gemeenten-ambtenaar-blijft-de-baas

VNO-NCW. (2018). Companies and scientists want national AI strategy. Retrieved August 6, 2019, from https://www.vno-ncw.nl/nieuws/bedrijven-en-wetenschappers-willen-nationale-ai-strategie?utm_source=e-mailnieuwsbrief&utm_medium=email&utm_campaign=AWTI+e-mail+alert

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, *2*(6), 6080. https://doi.org/10.1126/scirobotics.aan6080

Wang, W., & Siau, K. (2018). Artificial Intelligence: A Study on Governance, Policies, and Regulations. In *Thirteenth Midwest Association for Information Systems Conference* (Vol. 40, pp. 1–5). Saint Louis, Missouri. Retrieved from http://aisel.aisnet.org/mwais2018/40

Watts, S., & Stenner, P. (2005). Doing Q methodology: Theory, method and interpretation. *Qualitative Research in Psychology*, *2*, 67–91. https://doi.org/10.1191/1478088705qp022oa

Webler, T., Danielson, S., & Tuler, S. (2009). *Using Q method to reveal social perspectives in environmental research*. *SERI*. Greenfield MA. https://doi.org/10.1163/017353711X556989

Weller, A. (2017). Challenges for Transparency. In *ICML Workshop on Human Interpretability in Machine Learning* (pp. 55–62). Sydney. Retrieved from http://arxiv.org/abs/1708.01870

Winfield, A. F. T., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philosophical Transactions of the Royal Society A*, *376*(20180085), 1–13. Retrieved from http://eprints.uwe.ac.uk/37556 %0A

Yau, Y., & Lau, W. K. (2018). Big data approach as an institutional innovation to tackle Hong Kong's illegal subdivided unit problem. *Sustainability*, *10*, 2709–2726. https://doi.org/10.3390/su10082709

Zarsky, T. (2013). Transparent Predictions. *University of Illinois Law Review*, *2013*(4), 1503–1569.

Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology, & Human Values*, *41*(1), 118–132. https://doi.org/10.1177/0162243915605575

Zhang, B., & Dafoe, A. (2019). *Artificial Intelligence: American Attitudes and Trends*. Oxford

Appendices

Appendix 1: Search keywords (in English and Dutch)

English keyword	Dutch keyword
Artificial intelligence	Kunstmatige intelligentie
Algorithms	Algoritmen
Machine learning	Zelflerende machines
Deep learning	Deep learning
Explainability	Uitlegbaarheid
Interpretability	Interpreteerbaarheid,
Traceability	Traceerbaarheid, herleidbaarheid
Auditability	Auditbaar, toezicht, controleerbaarheid
Communication	Communicatie
Full transparency	Volledige transparantie
Limited transparency	Beperkte transparantie
Black-box	Zwarte doos
Opacity	Ondoorzichtigheid
Accountability	Verantwoordelijkheid
Legitimacy	Legitiem
Trust	Vertrouwen

Fairness	Eerlijkheid
Fear	Angst
Trade-offs	Afwegen
Privacy	Privacy
Perverse effects	Perverse

Appendix 2: Origin of statements - resulting from the keyword search

Type	Source
News platforms/forums	Tweakers Platform O ZIPconomy
Online newspapers	NRC Handelsblad Volkskrant Trouw
Online media	Omroepvereniging VPRO NPO Radio 1 Nederlandse Omroep Stichting (NOS)
Research institutes/think tanks	Rathenau Instituut Wetenschappelijke Raad voor het Regeringsbeleid Clingendael
Public sector	Ministerie van Justitie en Veiligheid Ministerie van Binnenlandse Zaken en Koninkrijksrelaties Tweede Kamer der Staten-Generaal Politie Nederland
Universities	Utrecht University Leiden University Tilburg University
Dutch private sector	KPMG Deloitte PwC Verdonck Klooster & Associates Paul Postma Marketing Consultancy
Consortiums/platforms	ECP Platform voor de InformatieSamenleving AINED Leer- en Expertisepunt Open Overheid

Political parties	VVD D66 SP

Appendix 3: Q-set i.e. list of 54 statements

#	Sub-category 1	Sub-category 2	Statement
9	Accountability		Transparency is important to get clarity about accountability; humans are and will remain responsible, and not 'the algorithm'.
14	Accountability		The question is, which measures are necessary to take responsibility: sometimes algorithmic transparency can be important for this. *
53	Accountability		It must be clear who decides: a human or the system
23	Auditability		The decision process of machine learning technology must be logged and stored. In such a manner that this information remains available for a minimum period to allow an inspection during audits.
37	Auditability	Limited transparency	The way the algorithm works, the implementation, and the data can be audited by judges and authorities, but not by civilians.
44	Auditability		The panel members preach for nuance about the role of the government surrounding transparency of algorithms like for the case for auditing. Many algorithms already have a form of supervision, for example in the financial sector. *
11	Black-box		Algorithms are for many people a new subject and often a 'black-box'. More transparency about this subject can therefore not hurt.
28	Black-box	Full transparency	"It may not be a black-box", said the Parliamentarian. That means not only government transparency but it also concerns the algorithms of companies.

32	Black-box		We should get used to some computer decisions being a black-box, just like human decisions.
4	Communication		The human-machine interface must be designed in such a way that a human always knows what the machine will do.
48	Communication		The transparency of a machine to the user at work has to increase. This can happen by means of communicating the current movement pattern and the possible diversions from that.
51	Communication		A machine with machine learning technology must be capable to give an adequate and fitting response to a human being.
1	Explainability		In many situations explainability remains needed: in healthcare for example, or when using an algorithm-produced analysis in a court-case.
19	Explainability	Trade-off (performance)	Without explainability of AI we cannot reach the potential of AI.
33	Explainability	Trade-off (performance)	Explainability can limit the added value of algorithms, because explainability places limits on the complexity that an algorithm can have.
25	Fairness: bias	Fear: risks/security	Algorithms can prevent discrimination, but they also carry the risk of being discriminatory.
31	Fairness: bias		Greater transparency can lead to discriminatory behavior.
42	Fairness: bias		Transparency is necessary to prevent that a 'bias' manifests in the machines.

21	Fairness: competitiveness		The art is to find a balance between algorithmic transparency on the one hand and the commercial interests of secrecy on the other hand.
30	Fairness: competitiveness		It should be prevented that transparency becomes a goal on its own, which would put the breaks on innovation.
47	Fairness: competitiveness		It is a fallacy that transparency of algorithms will put a break on innovation
5	Fear: impact		The demand for transparency will be big in some situations, because the workings of the algorithm can have a direct impact on people, such as with social security payments, monitoring and investigation, and case law.
40	Fear: impact	Limited transparency	The regulatory pressures are not always proportional, which for example is certainly the case for apps who work with less sensitive data. *
49	Fear: impact		It is only on the basis of transparency that we can protect the fundamental principles of our democracy.
3	Fear: risks/security		Many of the risks of AI happen as a result of a limited transparency of AI.
7	Fear: risks/security		The secret character of activities in the security domain strengthen the transparency paradox: civilians become more transparent for the government; while the algorithms of the government are barely transparent. This causes the balance of power between the civilians and state to shift.

17	Fear: risks/security		Disclosure of information about algorithms can be exempted in cases where it is highly necessary by law (including national security).
46	Full transparency		Full transparency for the case of algorithms used by AI is of little added value. To always make full transparency obligatory: of the source code of the algorithm and the data it uses, is not the solution.
15	Full transparency	Limited transparency	Next to full transparency there are some cases where more limited forms of transparency could be considered.
35	Full transparency		Algorithms used by the government need to be as transparent as possible.
26	Interpretability	Explainability	It is barely possible for AI to provide humans an interpretable explanation regarding the how and the why of its choices.
43	Interpretability	Communication	A machine with machine learning technology must declare which actions it will take and on the basis of which information.
54	Interpretability	Explainability	The choice of the algorithm depends on how well the algorithm can be interpreted and explained. *
2	Legitimacy		Legitimacy requires that government decisions are transparent, this is not evidently the case with governing algorithms. *
12	Legitimacy		Transparency can make an important contribution to the legitimate application of AI.

27	Legitimacy		Transparency is highly necessary to increase the legitimacy of the handling of governments.
29	Limited transparency		Only research that 'demonstrates' that the robot performs well will be made public. Because of this there will be a faulty reputation about the quality of the robot.
34	Limited transparency		Which parts should be made transparent depends on the specific needs.
52	Limited transparency		Transparency does not imply necessarily the insight in algorithms and data use. *
36	Perverse effects (information bombardment)		Technical transparency has its limits because the outputs of the algorithm are sometimes based on so many factors, that it is difficult to see which factors were decisive.
10	Perverse effects: (manipulation)		The problem of manipulation as a consequence of transparent data and models must not be overestimated, given that citizens in many circumstances cannot factually circumvent data collection. *
24	Perverse effects: (manipulation)		Companies and governments are also afraid that publicizing of algorithms will enable those who are malicious to manipulate the system. *
38	Privacy		Residents must maintain full control over their personal data.
45	Privacy		Limit the visibility of algorithms to prevent that the privacy of individuals can be affected.

50	Privacy		For organisations there can be privacy considerations to keep their algorithms and datasets secret.
6	Traceability		Data could be stored, with which experts, like developers of the algorithm can trace back to why something went wrong.
22	Traceability		It is not traceable whether a robot takes decisions on the basis of valid criteria (incorrect or non-scientific results are not valid).
39	Traceability		AI-systems must be traceable.
13	Trade-off (performance)	Fear: risks/security	Intelligence services cannot effectively operate when they have to be fully transparent.
16	Trade-off (performance)	Fear: impact	Losing effectiveness may, in a democratic society, not be a reason to drop the obligation for transparency. *
20	Trade-off (performance)		The way that AI systems are self-aware and achieve goals is often not transparent, and that is exactly why they are sometimes better than humans. *
8	Trust	Perverse effects (information bombardment)	Openness about the algorithm is difficult because users often do not understand how such an algorithm works. Such an explanation is too complicated and does not contribute to trust. *
18	Trust	Black box	An algorithm is for many a mysterious black box and that causes distrust.
41	Trust	Explainability	If the machine cannot give a better explanation than humans, then do not trust the machine.

Statements with a * were marked with an asterisk after the second pilot study because the participants found to be difficult to interpret.

Appendix 4: Q-sort correlation matrix

ID	2103	6501	3290	9171	5993	2981	9542	4610	2061	9916	6277	3573	7722	8442	6118	8108	8711	3722	7139	1592	6827	1402	5636	9276	4280	1312	5215	8573	7982	7514	3832
2103	100	17	8	12	23	13	26	32	13	37	30	33	29	28	20	36	21	17	16	-1	3	4	8	0	3	15	16	22	11	15	10
6501	17	100	3	-10	-5	-12	-22	-7	22	1	-10	3	6	-6	5	-8	-1	-23	14	36	19	27	-10	4	-2	-11	-46	20	9	35	25
3290	8	3	100	0	6	-13	44	30	30	21	32	30	37	14	23	6	-18	30	10	-17	22	-14	-18	-21	-32	-30	-5	4	-30	18	1
9171	12	-10	0	100	17	40	34	25	3	23	23	35	18	21	19	36	32	27	0	22	9	4	34	9	-6	-4	10	1	23	19	-16
5993	23	-5	6	17	100	29	34	37	23	27	30	14	25	24	17	24	9	34	18	-4	0	18	6	16	0	-2	22	23	2	-1	-7
2981	13	-12	-13	40	29	100	23	30	16	37	15	24	21	20	11	31	28	18	27	-9	4	-6	35	29	12	4	21	26	41	13	-5
9542	26	-22	44	34	34	23	100	59	29	60	63	46	53	40	12	49	8	63	-1	0	17	10	6	4	-4	4	23	15	-10	24	4
4610	32	-7	30	25	37	30	59	100	37	43	46	43	54	28	23	49	12	42	-2	-12	14	2	7	13	3	-2	1	9	-3	17	-18
2061	13	22	30	3	23	16	29	37	100	36	15	31	37	40	36	29	19	40	27	24	46	18	9	32	16	27	11	29	5	38	44
9916	37	1	21	23	27	37	60	43	36	100	54	48	52	50	31	33	23	45	6	0	22	-9	34	20	10	9	23	21	11	26	16
6277	30	-10	32	23	30	15	63	46	15	54	100	23	58	33	-10	27	2	47	-14	-13	-8	1	25	1	-20	-1	2	20	-10	21	-17
3573	33	3	30	35	14	24	46	43	31	48	23	100	27	25	33	25	36	26	11	9	19	-6	27	33	9	-3	10	38	26	35	9
7722	29	6	37	18	25	21	53	54	37	52	58	27	100	27	0	40	-2	46	2	-13	-1	-4	7	3	-18	-3	-4	9	-14	30	-7
8442	28	-6	14	21	24	20	40	28	40	50	33	25	27	100	30	38	45	56	7	20	28	6	23	36	9	28	19	17	14	31	38
6118	20	5	23	19	17	11	12	23	36	31	-10	33	0	30	100	28	15	22	28	9	26	1	6	7	12	11	7	-8	0	21	20
8108	36	-8	6	36	24	31	49	49	29	33	27	25	40	38	28	100	11	40	6	7	-9	19	7	12	19	23	4	14	1	20	7
8711	21	-1	-18	32	9	28	8	12	19	23	2	36	-2	45	15	11	100	25	25	37	26	5	39	36	17	25	29	30	50	14	31
3722	17	-23	30	27	34	18	63	42	40	45	47	26	46	56	22	40	25	100	7	18	5	9	3	12	9	11	36	22	-9	9	22
7139	16	14	10	0	18	27	-1	-2	27	6	-14	11	2	7	28	6	25	7	100	8	10	-4	27	36	7	9	24	0	17	14	25
1592	-1	36	-17	22	-4	-9	0	-12	24	0	-13	9	-13	20	9	7	37	18	8	100	29	35	-1	19	27	24	-3	17	30	43	47
6827	3	19	22	9	0	4	17	14	46	22	-8	19	-1	28	26	-9	26	5	10	29	100	11	11	16	7	17	5	-1	17	37	36
1402	4	27	-14	4	18	-6	10	2	18	-9	1	-6	-4	6	1	19	5	9	-4	35	11	100	-29	6	16	36	1	24	7	28	31
5636	8	-10	-18	34	6	35	6	7	9	34	25	27	7	23	6	7	39	3	27	-1	11	-29	100	44	18	19	20	9	47	5	-12
9276	0	4	-21	9	16	29	4	13	32	20	1	33	3	36	7	12	36	12	36	19	16	6	44	100	37	31	22	34	40	29	28
4280	3	-2	-32	-6	0	12	-4	3	16	10	-20	9	-18	9	12	19	17	9	7	27	7	16	18	37	100	29	23	14	25	-7	33
1312	15	-11	-30	-4	-2	4	4	-2	27	9	-1	-3	-3	28	11	23	25	11	9	24	17	36	19	31	29	100	16	-1	30	14	28
5215	16	-46	-5	10	22	21	23	1	11	23	2	10	-4	19	7	4	29	36	24	-3	5	1	20	22	23	16	100	19	13	3	29
8573	22	20	4	1	23	26	15	9	29	21	20	38	9	17	-8	14	30	22	0	17	-1	24	9	34	14	-1	19	100	17	7	21
7982	11	9	-30	23	2	41	-10	-3	5	11	-10	26	-14	14	0	1	50	-9	17	30	17	7	47	40	25	30	13	17	100	18	17
7514	15	35	18	19	-1	13	24	17	38	26	21	35	30	31	21	20	14	9	14	43	37	28	5	29	-7	14	3	7	18	100	38
3832	10	25	1	-16	-7	-5	4	-18	44	16	-17	9	-7	38	20	7	31	22	25	47	36	31	-12	28	33	28	29	21	17	38	100

Appendix 5: Flagged Q-sorts: defining sorts indicated by an X

Q-sort ID	Factor Loadings		
	1	2	3
2103	**0.4332X**	0.1249	0.1152
6501	-0.1341	-0.1775	**0.5853X**
3290	**0.4996X**	-0.4984	0.1348
9171	0.4033	0.3288	-0.1159
5993	**0.4676X**	0.1350	-0.0443
2981	0.3682	0.5419	-0.1864
9542	**0.8387X**	-0.0481	0.0280
4610	**0.7470X**	-0.0201	-0.0285
2061	0.4471	0.1002	0.6000
9916	**0.7275X**	0.2196	0.0881
6277	**0.7321X**	-0.1069	-0.1628
3573	0.5493	0.2703	0.1563
7722	**0.7411X**	-0.1943	0.0037
8442	0.5296	0.3100	0.3250
6118	0.2866	0.1061	0.3188
8108	**0.5780X**	0.1409	0.0934
8711	0.1591	**0.6733X**	0.2191

3722	**0.6916X**	0.0897	0.1327
7139	0.0606	0.3484	0.2010
1592	-0.1025	0.2479	**0.6665X**
6827	0.1244	0.1100	**0.5496X**
1402	-0.0209	-0.0125	**0.5318X**
5636	0.1857	**0.6787X**	-0.2260
9276	0.1128	**0.6527X**	0.2404
4280	-0.1068	**0.4919X**	0.2131
1312	-0.0194	0.4194	0.3152
5215	0.1891	**0.4585X**	-0.0610
8573	0.2392	0.2664	0.2162
7982	-0.0814	**0.7289X**	0.1095
7514	0.2952	0.0508	0.6059
3832	-0.0345	0.2199	**0.7628X**
% Explained variance	**19**	**12**	**11**

Appendix 6: Factor distinguishing statements

Statement	Factor 1		Factor 2		Factor 3	
	Q-SV	Z-SCR	Q-SV	Z-SCR	Q-SV	Z-SCR
1	4	1.56*	3	1.06	1	0.51
2	1	0.54	-4	-1.28**	1	0.18
3	1	0.45**	-5	-1.55**	-1	-0.38**
4	0	0.27**	-4	-1.12	-4	-0.99
5	4	1.45	2	0.68*	4	1.27
6	-1	-0.07**	4	1.37	5	1.77
7	5	1.59**	1	0.63**	-1	-0.35**
8	-5	-1.79**	1	0.21**	5	1.66**
9	5	1.58**	2	0.71	3	0.83
10	-2	-0.55	-1	-0.61	-3	-10.85
11	3	0.73	2	0.67	-1	-0.44**
12	5	1.62**	0	0.20	0	-0.03
13	-1	-0.07	5	2.18**	-2	-0.50
14	0	0.09	0	0.02	-1	-0.41
15	-1	-0.46**	5	1.43*	2	0.80*
16	5	1.58**	-1	-0.59	0	-0.12
17	0	-0.05	4	1.35**	1	0.46
18	1	0.49	1	0.37	-4	-1.17**
19	3	0.74**	-2	-0.82**	-5	-1.79**
20	-5	-1.86*	-4	-1.25*	5	2.36**
21	-2	-0.55	5	1.49**	-2	-0.75
22	-3	-0.95	-5	-1.85*	-4	-1.11
23	1	0.35**	5	1.47	4	1.39
24	-1	-0.32**	4	1.28	4	1.36
25	4	1.09**	0	-0.16	0	-0.25
26	-2	-0.60	-2	-0.63	2	0.59**
27	4	1.49*	-1	-0.34**	3	0.87*
28	1	0.34	1	0.2	1	-0.03
29	-1	-0.25**	-5	-1.67	-4	-1.33
30	-3	-1.05	1	0.24**	-2	-0.71
31	-2	-0.69	-4	-1.27*	0	-0.27
32	-5	-1.81**	-3	-0.96**	2	0.65**
33	-5	-1.81**	-1	-0.53**	4	1.32**
34	0	-0.03**	3	1.08	2	0.70
35	3	0.87	-1	-0.57**	3	1.00
36	-4	-1.52**	-1	-0.41	0	-0.15
37	-3	-1.10**	3	1.06**	0	-0.25**

38	-1	-0.12**	-3	-1.00**	5	1.73**
39	0	0.01**	4	1.32	2	0.78
40	2	0.60	-2	-0.79**	3	0.86
41	-2	-0.52**	-5	-1.31	-5	-1.44
42	2	0.61**	-3	-1.03	-3	-0.97
43	2	0.69**	0	0.03**	-3	-0.87**
44	-4	-1.31	-2	-0.82	-3	-0.97
45	-4	-1.40*	0	0.13**	-3	-0.85*
46	-3	-1.29**	3	1.11**	-2	-0.50**
47	3	0.82**	0	-0.01	-1	-0.48
48	2	0.64**	-2	-0.74	-2	-0.73
49	2	0.72*	-3	-0.96**	1	0.19*
50	-3	-0.70	2	0.80**	-1	-0.36
51	0	0.11	2	0.79*	1	0.07
52	-4	-1.30	1	0.45**	-5	-1.74
53	3	0,86	3	0,88	3	1,14
54	1	0.32**	-3	-0.93*	-5	-1.68*

The factor distinguishing statements are based on both Q-sort value (Q-SV), Z-scores (Z-SCR), and P-values. White is $p > 0.05$; Light blue (*) is $p < 0.05$; Dark blue (**) is $p < 0.01$; Red text implies a negative z-score, Black text implies a positive z-score. Q-SV are on a scale from -5 to 5, where 5 is “most how I think” and -5 “least how I think”

Appendix 7: Demographics of high loadings per factor

Factor 1 demographics

ID	Loading	AI knowledge	Algorithm knowledge	Employment	Education	Gender
2103	0.4332X	3	1	Private	Vocational	Male
3290	0.4996X	3	3	Private	Vocational	Female
5993	0.4676X	4	3	Public	Bachelor	Male
9542	0.8387X	3	3	NP/R/A	Master	Female
4610	0.7470X	3	1	Public	Vocational	Male
9916	0.7275X	2	2	NP/R/A	Master	Female
6277	0.7321X	6	5	Public	Master	Male
7722	0.7411X	1	1	NP/R/A	Vocational	Male
8108	0.5780X	1	1	Private	Vocational	Female
3722	0.6916X	4	4	NP/R/A	PhD	Male

Factor 2 demographics

ID	Loading	AI knowledge	Algorithm knowledge	Employment	Education	Gender
8711	0.6733X	4	1	NP/R/A	PhD	Male
5636	0.6787X	4	3	Private	Bachelor	Male
9276	0.6527X	4	4	Public	Master	Male
4280	0.4919X	6	6	NP/R/A	PhD	Male
5215	0.4585X	7	7	Private	PhD	Male

7982	0.7289X	4	4	NP/R/A	Bachelor	Female

Factor 3 demographics

ID	Loading	AI knowledge	Algorithm knowledge	Employment	Education	Gender
6501	0.5853X	4	3	Private	Vocational	Male
1592	0.6665X	2	2	NP/R/A	PhD	Male
6827	0.5496X	3	5	NP/R/A	PhD	Female
1402	0.5318X	7	7	Private	High School	Male
3832	0.7628X	3	5	Student	Master	Female

Appendix 8: Questionnaire (in Dutch)

Vragenlijst

Datum: _____

Uw leeftijd: _____

Uw huidige/hoogst genoten opleiding: _____

Uw geslacht (graag aanvinken)
- o Vrouw
- o Man
- o Anders**:**
- o Geef ik liever niet aan

Uw woonomgeving (graag aanvinken):
- o Stedelijk
- o Landelijk

Uw werk of meest recente beroep (graag aanvinken):
- o De publieke sector
- o De private sector
- o De non-profit-, onderzoeks-, of academische sector
- o Ik ben een student

Hoeveel kennis heeft u van (graag invullen, op schaal van 1 tot 7):
- o Algoritmen: _____
- o Kunstmatige intelligentie: _____

Ter indicatie: 1 = absoluut geen kennis, 4 = neutraal/kennis, 7 = heel veel kennis

Kunt u (graag aanvinken):
- o Internetbankieren
- o Online winkelen en afrekenen
- o Gebruikmaken van een zoekmachine
- o E-mailen

ID-nummer: _____ (in te vullen door de onderzoeker)

Appendix 8: questionnaire (in Dutch, p. 2/2)

Waar ligt het middelpunt van de stellingen (graag omcirkelen):

-5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5

Zijn er stellingen waar u een toelichting over wilt geven?

Hebt u toelichtingen over uw sorteeruitkomst (of niet genoeg ruimte hierboven)?

ID-nummer: _____ (in te vullen door de onderzoeker)

Appendix 9: Consent form (in Dutch)

Toestemmingsformulier

Beste_____,

Hierbij verwelkom ik uw deelname aan mijn masteronderzoek genaamd: *"Transparency of Artificial Intelligence: A Discourse Analysis of Dutch Public Opinion Using Q-Methodology".*Het doel van deze studie is om een beter beeld te krijgen over de publieke meningen in Nederland over de transparantie van Kunstmatige Intelligentie. Dit onderzoek wordt geleid door mijzelf, Stefano Sarris, een masterstudent Public Administration aan de Universiteit Leiden.

Ik zal u vragen om een rangschikking te doen, gebaseerd op uw mening, over stellingen die gaan over de transparantie van Kunstmatige Intelligentie. Deze onderzoeksmethode heet de Q-methodologie. Tijdens het interview zult u gedetailleerde informatie ontvangen over deze methodologie. De totale duur van het interview is naar verwachting 30 minuten.

Deelname aan deze studie is 100% op vrijwillige basis --"u" (de participant) mag uw deelname aan het interview op elk moment stopzetten, u hoeft hiervoor geen uitleg te geven. U zult geen financiële compensatie of andere vormen van vergoedingen ontvangen voor uw deelname.

De identiteiten van de participanten worden volledig confidentieel gehouden. Uw persoonlijke informatie zal niet gedeeld worden met personen buiten het onderzoeksteam. De data die gebruikt zal worden buiten het onderzoeksteam zal vrij zijn van alle identificerende informatie. U mag uw toestemming voor het gebruik van uw data op elk moment intrekken, zonder dat u hiervoor een verklaring hoeft af te leggen.

Gelieve graag contact met mij op te nemen mocht u nog enige vragen of opmerkingen hebben betreft uw deelname aan het interview. Ik zal uw vragen met genoegen beantwoorden. Ik ben vrijblijvend te bereiken via (e-mail onderzoeker).

Om toestemming te geven moet u het volgende ondertekenen:
Ik, _____ (de participant), verklaar hierbij dat ik het toestemmingsformulier heb gelezen en dat ik instem met de rechten en plichten zoals beschreven. Ik ben op de hoogte van mijn rol en mijn rechten in het onderzoek. Ik neem vrijwillig deel aan het onderzoek en zal naar juistheid de vragen beantwoorden. Ik heb een kopie van deze brief meegekregen.

De participant		***De onderzoeker***	
Naam	**_____**	**Naam**	**_____**
Datum	**_____**	**Datum**	**_____**

Handtekening _____ **Handtekening** _____