NTNU
Kunnskap for en bedre verden

Department of Industrial Economics and Technology Management

TIØ4562 - Strategy, Innovation and International Business Development, Specialization Project

The *how* and *why* of responsible artificial intelligence

A systematic literature review

Author:
Andreas Bjelland Berg

Supervisor:
Nhien Nguyen

December, 2022

Abstract

Artificial intelligence (AI) is everywhere, and has great potential for doing good, through its autonomous and scalable nature. However, several cases have shown that this nature also brings with it a potential for negative results, which, due to the vast reach of AI technology, can have large consequences for society. To avoid these negative results, AI systems should be developed to be responsible.

This paper reviews 54 papers on responsible AI, with the goal of creating an understanding of how AI systems can be designed to be responsible. This review results in a set of four core principles – Autonomy, Beneficence, Non-maleficence and Justice – that create responsible value by themselves, as well as seven instrumental principles for responsible AI – Transparency, Accountability, Trust, Sustainability, Privacy and Others – that create responsible value by facilitating for the core principles. Additionally, the reviewed literature included three clusters of antecedents for responsible AI, as well as three clusters of business advantages that can be gained by adopting responsible AI practices.

The review also uncovered a gap between abstract principles and actual responsible AI development. To help bridge this, this paper introduces a project-level framework for responsible AI development, the EPNIS framework. This framework is compared to existing frameworks for responsible AI, to fit it into the current ecosystem of responsible AI methods. Finally, the results of the review are used to point out future research that is needed within the field of responsible AI

Oppsummering

Kunstig intelligens (KI) er over alt, og har stort potensiale for gode resultater, via dets autonome og skalerbare natur. Flere hendelser har likevel vist at denne naturen også har potensiale for negative resultater, noe som, grunnet den store rekkevidden til kunstig intelligens, can ha store konsekvenser for samfunnet. For å unngå disse negative resultatene bør KI-systemer utvikles på en ansvarlig (engelsk: *responsible*) måte.

Denne oppgaven analyserer 54 dokumenter om ansvarlig kunstig intelligens, med mål om å skape en forståelse for hvordan KI-systemer kan utvikles på en ansvarlig måte. Analysen resulterer i fire kjerneprinsipper for ansvarlig kunstig intelligens – Autonomi (*Autonomy*), Velgjørighet (*Beneficence*), Ikke-skade (*Non-maleficence*) og Rettferdighet (*Justice*) – som skaper ansvarsverdi på egenhånd, samt syv instrumentelle prinsipper – Åpenhet (*Transparency*), Ansvarlighet (*Accountability*), Pålitelighet (*Trust*), Bærekraftighet (*Sustainability*), Rett til privatliv (*Privacy*) og Andre (*Others*) – som skaper verdi gjennom å støtte og fasilitere for kjerneprinsippene. I tillegg finner analysen tre grupper med årsaker (*antecedents*) til at selskaper utvikler ansvarlig kunstig intelligens, samt tre grupper med fordeler (*business advantages*) bedrifter kan oppnå ved å bruke metoder for ansvarlig kunstig intelligens.

Analysen oppdaget også en avstand mellom abstrakte prinsipper og faktisk implementasjon av ansvarlig kunstig intelligens. For å minimere denne avstanden introduserer dette papiret et rammeverk for ansvarlig KI-utvikling på prosjektnivå, EPNIS-rammeverket. Dette sammenlignes med eksisterende rammeverk for ansvarlig kunstig intelligens, for å finne dets plass i det nåværende økosystemet for ansvarlig kunstig intelligens. Avslutningsvis brukes resultatene av analysen til å peke ut fremtidig forskning som behøves innenfor ansvarlig KI-feltet.

Table of Contents

List of Figures

List of Tables

1 Introduction

Artificial intelligence (AI) is everywhere. It is used to recommend products on Amazon, prevent abusive comments and messages on Facebook, and find interesting profiles on Twitter (Cooper, 2022). AI helps users find the websites they are looking for when searching on Google (Raghavan, 2020), and the movies they want to watch on Netflix (Netflix Research, n.d.).

Outside of the internet, AI is used in applications ranging from detecting faults in power supply networks and forecasting energy usage (Quest et al., 2022), to predicting traffic (Sarker, 2021) and making tactical decisions in sport competitions (Ding, 2019). AI has potential to make cars safer (Myers, 2022) and self-driving (BMW, 2020), save lives using drones (AI for Good, n.d.), and fly planes autonomously (Airbus, n.d.). In the domain of healthcare, AI is used for applications such as improving drug safety (Basile et al., 2019), identifying diseases (Kermany et al., 2018) and cancer (Hosny et al., 2018) using images, battling pandemics such as COVID-19 (Vaishya et al., 2020), and restoring movement to patients with quadriplegia (i.e., loss of function in limbs) (Jiang et al., 2017). All this is done automatically, with minimal human intervention.

This, naturally, has enormous potential. PwC (2017) estimates that AI has potential to contribute 15.7 trillion USD by the year 2030, while Chui et al. (2018) place the potential impact at between 9.5 and 15.4 trillion USD annually. Organisations surveyed by Marshall and Goehring (2020) report an average increase in revenue of 6.3 percentage points, directly caused by adopting AI technologies. The potential of AI is not only through economic results, however. By minimizing human intervention, AI technologies enable quick scaling of business processes, ensuring maximum efficiency is reached (Lardi, 2021). At the same time, AI empowers individual employees, increasing their competence and job satisfaction (Ransbotham et al., 2022). The autonomous and scalable nature of AI opens up a sea of new possibilities, such as detection of disease before any symptoms show up (Agrawal, 2021).

However, such an explosion in artificial intelligence comes with significant drawbacks. AI systems may discriminate against certain groups of individuals, place human lives at risk, or increase polarization of society (Mikalef et al., 2022). Additionally, systems like autonomous weapons may use AI for intentionally causing harm (Eitel-Porter, 2021). Several negative consequences of AI have already occurred. Amazon's AI system for recruitment, designed to automatically filter resumes sent to the company, was shut down after it was discovered that it greatly preferred male applicants (Dastin, 2018). An AI system used by US courts to pre-

dict recidivism, called "Correctional Offender Management Profiling for Alternative Sanctions" (COMPAS), falsely predicted a high risk for recidivism for black people almost twice as often as for white people (Angwin et al., 2016). In a similar fashion, Google Photo's automatic picture-tagging algorithm, designed to simplify organizing pictures in the application, labeled a picture of two black people as "gorillas" (Barr, 2015).

The risks of wrongful use of AI, as illustrated by these cases, show the importance of designing and developing AI in an ethically responsible way. While the debate of AI ethics has been ongoing almost since the term AI was first used (see e.g. Samuel, 1960; Wiener, 1960), responsible AI is a recent term, used to describe AI systems developed with potential consequences for humanity in mind. Current work in the field of responsible AI largely revolves around creating sets of ethical principles to follow when developing systems, and little work has been done on ways to implement these principles in the actual systems and processes (Barredo Arrieta et al., 2020).

Based on this background, the goal of this paper is to create an understanding of how an AI system can be designed to be considered responsible. In order to achieve this goal, it is also necessary to understand why organisations should spend the resources it takes to ensure responsibility. As such, the goal is broken into three separate research questions (RQs) – RQ1, RQ2, and RQ3 – presented below.

RQ1 What are the criteria for an AI system to be considered responsible?
RQ2 What are the antecedents for responsible AI adoption?
RQ3 What business advantages can be achieved by choosing responsible AI systems over other AI systems?

To achieve its goal, this paper conducts a systematic literature review of the academic field of responsible AI. The goal of this review is to summarize the standing of the field, creating a "snapshot" of the current state of research. This paper then uses the finding of this review to create a framework, designed to concretize and simplify the process from finding a set of principle suitable for a system and context, and converting these to actual implementations within the system.

There have been some attempts at compiling the work done on responsible AI by reviewing academic literature on the subject. Anagnostou et al. (2022) conducts a literature review with the goal of finding common principles, and how they apply, for a specific set of industries. Similarly, Lukkien et al. (2021) and Siala and Wang (2022) perform a literature reviews specifically focused on responsible AI in healthcare. The review conducted by Barredo Arrieta et al. (2020) targets explainable AI, rather than responsible AI. In a similar fashion, the goal of Morley et al. (2020) is to discover technical tools for implementing responsible AI, while Merhi (2022) aims

to discover barriers preventing responsible AI from being created. Finally, Y. Wang et al. (2020) aims to find the benefits of implementing responsible AI practices. As such, this paper presents novel knowledge, in that it is the first to conduct a systematic literature review of academic literature with the goal of discovering how AI systems can be made responsible from a non-technical perspective.

This paper is structured as follows. Section 2 discusses some background information that is relevant for understanding the rest of the paper, before Section 3 presents the methodology used for the review. The findings of the review are given in Section 4. The section starts off with descriptive results, before some key concepts are defined in Section 4.2. Section 4.3, 4.4 and 4.5 aims to answer RQ1, RQ2 and RQ3, respectively. A discussion of the findings is presented in Section 5, with Section 5.2 presenting the new framework. Finally, the paper wraps up with a conclusion in Section 6.

2 Theoretical background

In order to understand the rest of the paper, there are some concepts that are important to know about. First, digital technology and AI are introduced at a conceptual level. The notion of stakeholders is discussed, and how they can be included in the design and development of projects.Due to the connections between responsible AI and responsible innovation, a brief introduction to Responsible Research and Innovation is given, before the section rounds out by giving a basic introduction to ethics, mostly focused on applied ethic.

2.1 Digital technology

Digital technology is the use of software, hardware and networking technologies, i.e. computers, to digitally process information (Weisha, 2021). Such digitization has the possibility of connecting people and organisations across wide geographical distances (Forman & van Zeebroeck, 2019), and has impacted a wide range of services, ranging from diabetes awareness campaigns (O'Mara et al., 2012) to energy sustainability (J. Wang et al., 2022) and art education (W. Wang, 2020).

As digital technology has been shown to increase organisational innovation (Weisha, 2021), efficiency (Grober & Grober, 2020) and revenue (Mohapatra et al., 2022), a wide range of organisations have adopted and applied such technologies to their business. To successfully integrate digital technology into an organisation often require organisational changes, a process that is commonly referred to as digital

transformation (Kretschmer & Khashabi, 2020).

Digital transformation is built on five building blocks, as described by Ross et al. (2018). First, organisations looking to adapt digital technology need to have an operational backbone, a set of standardized systems and processes that lets digital technology cooperate, thus increasing efficiency of the organisation. Secondly, organisations need a digital platform, a set of reusable components, facilitating rapid development and experimentation. To support their network, organisations should create an external developer platform, providing easy access for external partners to access and contribute to the organisations' digital technology. In order to get value from these digital platforms, organisations should ensure they have shared customer insights, i.e., a knowledge of what customers are willing to pay for, and an accountability framework, to minimize hierarchy and facilitate rapid delivery of digital technology.

Recent developments of digital technology have led to rapid improvements in the efficiency and flexibility of digital systems, leading some researchers to refer to the current state of digital technology as Industry 4.0 – the Fourth Industrial Revolution (Lasi et al., 2014). This term covers a wide range of modern digital technology, mostly revolving around automation and digitization (Y. Lu, 2017), including technology such as cloud computing, i.e., off-site servers and computers available for rent; the Internet of Things, i.e., the use of distributed and interconnected sensors, tools and machines that can collect data and communicate with each other to automate and optimize processes (Xu et al., 2018); cyber-physical systems, i.e., the interconnectedness of physical machines and digital systems, giving digital technology a way to interact with a physical environment (Lasi et al., 2014); mobile computing, i.e., smart phones and portable devices; big data, i.e., the use of large-scale collection and storage of data; and artificial intelligence (Y. Lu, 2017).

2.1.1 Artificial intelligence

Artificial intelligence has been referred to as "the next evolution of digital transformation" (Graham, 2020). Where standard digital technology has been used for diabetes awareness, energy sustainability and art education, AI has been used to predict the risk of developing diabetes (Ellahham, 2020), has been named an enabler of 134 of the United Nations Sustainable Development Goals (Vinuesa et al., 2020), and has been used to make systems capable of autonomously creating art (Cetinic & She, 2022).

As AI is a form of digital transformation, the same building blocks, described by Ross

et al. (2018), must be in place to succeed with AI projects, with some adaptations. As AI learns from data (Graham, 2020), organisations looking to successfully adopt AI technologies must ensure their operational backbone and digital platforms are designed to standardize and systematize data collection and storage. This way, AI technologies can be developed in the same rapid fashion as other digital technology, without having to repeat the process of data collection each time (Werder et al., 2022). Additionally, in scenarios where such data may contain personal information, organisations should expand the notion of shared customer insights. Where ordinary digital transformation only needs insight into what digital services a customer is willing to pay for, organisations looking to adopt AI technologies should also consider how the process of collecting and using data, as well as the data itself, may affect customers and noncustomers alike (Dignum, 2021).

Artificial intelligence is further explored and defined in Section 4.2.1.

2.2 Stakeholders

The notion of stakeholders, whose fame is often attributed to Freeman and Reed; Freeman (1983, 1984; see, e.g., Mitchell et al., 1997), is that companies have obligations for other groups than only those holding shares in the company (Freeman & Reed, 1983). When making decisions, organisations should therefore not only consider what impact a decision may have on the economy of the organisation, but also what effect it may have on other affected people (Freeman & Reed, 1983). These term stakeholders, then, is used to represent these other affected people.

Mitchell et al. (1997) note that there have been multiple, somewhat contradictory definitions of stakeholders throughout history, and that these mostly differ along two axes. The first axis is that of broadness, where definitions range from narrow ones, such as "Any identifiable group or individual on which the organisation is dependent for its continued survival" (Freeman & Reed, 1983), to broad definitions like "any group or individual who can affect or is affected by the achievement of the organization's objectives" (Freeman, 1984; as cited in Mitchell et al., 1997, p. 856). The second axis is that of primary versus secondary stakeholders, where primary stakeholders are typically defined as those who bear risk through commitment of resources, e.g., shareholders, employees, customers, suppliers and the environment (Hillman & Keim, 2001), while secondary stakeholders are those who can influenced and are influenced by an organisation, but that the company can survive without (Benn et al., 2016). As it is outside of the scope of this paper to conduct a deep-dive into definitions of stakeholders, interested readers are guided to see the work

of Mitchell et al. (1997).

Who should be considered as stakeholders vary based on the context and nature of a project, but there exists several universal tools for identifying these. Reed et al. (2009) propose three methods for identifying stakeholders. First, organisations may use domain experts to provide stakeholders based on their experiences with similar projects. This method requires few resources, but has the greatest potential for bias among the methods. Secondly, organisations may conduct focus group sessions, where a small group of internal and external parties brainstorm potential stakeholders for a project. Finally, once some stakeholders have been identified, organisations may use snowball sampling, i.e., interviewing already identified stakeholders, to identify new categories. These methods can be used alongside each other to the number of relevant stakeholders that are identified. Once all potential stakeholders have been identified, organisations may use stakeholder maps to categorize them, giving a visual overview of all stakeholders, their importance for the project and what communication strategies are most likely to give good results for each stakeholder group (Hoory & Bottorff, 2022).

2.3 Responsible Research and Innovation

Responsible Research and Innovation (RRI) is a process of innovating based not only on what provides the largest economic value, but combining economic gain with ethical values and societal needs to develop products and services that benefit society as a whole (Werker, 2020). The term was first used in 2013, but has gained popularity after its inclusion in the European Commission's Horizon 2020 Framework Program (Burget et al., 2017).

Two core methods exist to ensure research and innovation processes are conducted in a responsible way. First, Werker (2020) highlights the need to include stakeholders, as they are needed to fully understand the effects of research processes – and potential alternatives that can be pursued. The definition of stakeholders should be broad, to ensure as many people as possible are included, and the collaboration should be an ongoing process, where the stakeholders actively take part in the research project throughout its life cycle (Schomberg, 2013).

Secondly, responsible research projects should be developed according to a set of conceptual dimensions (Burget et al., 2017). Although many different dimensions exist, Burget et al. (2017) finds that the dimensions Anticipation, i.e., the ability to envision the future of research, and how work can change the future; Inclusion, i.e.,

the early inclusion of stakeholders; Responsiveness, i.e., the ability to respond to risk and act in a transparent way; Reflexivity, i.e., the ability to reflect on limitations of research, and how it may affect people with different backgrounds; Sustainability, i.e., ensuring resource-efficiency is increased with new innovation, and doing so in a sustainable way; and Care, i.e., the ability to emphasize, and get stakeholders to care about responsible research, are frequently discussed in RRI literature, and present a good framework for ensuring research is conducted in a responsible way.

2.4 Ethics

Ethics revolve around questions of what is right and wrong, and what actions is "correct" for a given setting. The field is typically divided into three branches – meta-ethics, which looks at the language and sources of ethics; normative ethics, which looks at norms and standards for action; and applied ethics, which looks at how these norms and standards can be implemented in given contexts (Clarke, 2019). As the field of responsible AI looks at both norms and standards for AI development and how it can be applied to real development scenarios, the research being done on responsible AI falls under both normative and applied ethics.

As the ethics is a massive field of research, delving too deep into ethical theory is outside the scope of this paper. Still, this surface-level introduction should be enough to understand the ethics terms used throughout the paper.

3 Method

In order to have a factual basis for the literature review, it was conducted following the methodology laid out by Tranfield et al. (2003). This consists of three stages, for a total of 10 phases. An overview of these phases is shown in Table 1.

3.1 Stage I – Planning the review

According to Tranfield et al. (2003), the goal of stage I is to form a review panel – whose job it is to direct the process and review disputes, and develop a concrete plan for the review. This stage also contains any scoping reviews conducted to get an overview of the selected topic (Tranfield et al., 2003).

Table 1: Overview of the methodology used for this literature review. From Tranfield et al. (2003).

Stage I	**Planning the review**
Phase 0	Identification for the need for a review
Phase 1	Preparation of a proposal for a review
Phase 2	Development of a review protocol
Stage II	**Conducting a review**
Phase 3	Identification of research
Phase 4	Selection of studies
Phase 5	Study quality assessment
Phase 6	Data extraction and monitoring progress
Phase 7	Data synthesis
Stage III	**Reporting and dissemination**
Phase 8	The report and recommendations
Phase 9	Getting evidence into practice

Identification for the need for a review The identification for the need for a review is shown in Section 1. The review panel for this paper consist of a single person – the supervisor assigned to the paper. Any confusions were solved with her help, and she also directed the process and made sure it progressed over time, thus fulfilling the role of a review panel as described by (Tranfield et al., 2003).

Preparation of a proposal for a review As the author has limited experience with the field of responsible AI, a scoping study was conducted ahead of time. with the goal of getting an overview of the field of responsible AI, in order to help shape the main study conducted after. To avoid having to filter out technical articles, and thus save time, the scoping study primarily focused on papers written from a business perspectice.Thus, the scoping study was done searching two academic databases – Scopus and Web of Science – for papers containing the strings "Responsible AI" or a combination of the word "Responsible" with one of "Machine learning", "Artificial intelligence" or "AI", that also included the word "Business*". The complete search strings used for the scoping study is included in Appendix A.1.

Development of a review protocol Based on the findings of the scoping study, an informal review protocol was created. Tranfield et al. (2003) state that the review protocol ideally should be a formal document, and should be registered with a third-party group. While the review protocol for this paper was an informal document, and only kept internally, it was still developed before conducting the study, and not

changed after the review had started. This is in line with other review protocols for management studies (Tranfield et al., 2003), which this should be considered as. The review protocol consisted of research questions (presented in Section 1), the steps of the study (presented later in this section), and the eligibility criteria (shown in Table 2).

3.2 Stage II – Conducting the review

The goal of this stage is to actually conduct the review, following the steps laid out in the review protocol (Tranfield et al., 2003). The review should consist of "a comprehensive, unbiased search" (Tranfield et al., 2003, p. 215), a filtering based on the predefined eligibility criteria, and comprehensive coding of the results.

Identification of research The main finding from the scoping review was that the selection of search strings were too wide – allowing all combinations of "Responsible" and synonyms for AI, while also too narrow – requiring the inclusion of the word "Business*" in the results. The strings also lead to many papers that were not relevant for the selected research questions. Based on these findings, it was concluded that the literature review should simply search for papers including the terms "Responsible AI" or "Responsible Artificial Intelligence". Tranfield et al. (2003) argue that searches should be conducted for both published papers, conference proceedings, unpublished papers, as well as several non-academic sources. In order to limit the scope of this paper, searches were conducted for published articles, conference papers, editorials and book chapters. The complete search strings that were used are included in Appendix A.2.

The same two databases were used as in the scoping study – Scopus and Web of Science. They were both searched on October 4, 2022. The output of the search was a list of 397 academic papers.

Selection of studies The selection of studies was done in five steps. Although Tranfield et al. (2003) suggest conducting the study selection with more than one researcher, this was done by the author alone, due to the nature of this project. Step 3 – abstract screening – was adapted to minimize the bias that could come as a result of this. The steps of the selection were as follows:

(1) **Duplicate removal** After searching the two databases, two overlapping sets of papers were retrieved. When combining these, any duplicates between them

were removed. This step removed 108 duplicates from the dataset.

(2) **Title screening** In order to quickly filter out irrelevant papers, the titles of the results were screened, to check whether they fit with the eligibility criteria. Any records where there was uncertainty on whether the paper should be accepted were included and used in the next step. This step removed 72 records from the dataset, due to not fulfilling the eligibility criteria.

(3) **Abstract screening** The abstracts of the included results were screened, to check whether they fit with the eligibility criteria. As the review was conducted by one person, and the eligibility assessment is subjective, this may lead to bias in the results (Tranfield et al., 2003). To limit this, and ensure that all relevant papers were included in the review, this step was done twice.

   (3a) **Round one** First, all abstracts of the databased were reviewed. In cases where there was uncertainty, i.e. cases where it was unclear whether the paper should be included in the review, these were marked as uncertain and used in round two. This step removed 58 records from the dataset.

   (3b) **Round two** The abstracts of the uncertain cases from round one (3a) were reviewed again. This assessment was conducted on a new day, in order to ensure that the papers got a fair second chance. The cases that were still uncertain were excluded from the review. This step removed 102 records from the dataset.

(4) **Full-text screening** All papers accepted thus far had their full text assessed, to check whether the papers fit with the eligibility criteria. Any uncertain cases in this step were supposed to be excluded. No cases were marked as uncertain during the actual review. This step removed 20 papers from the dataset.

(5) **Snowballing** The reference lists of all accepted papers were scanned, to find other relevant sources. In order to only include relevant papers, all possibly relevant sources went through the same steps as mentioned here. Any papers that were included from snowballing then had their reference lists scanned, until no more relevant sources were found. Snowballing was done to ensure maximum inclusion of relevant documents, and has been recommended by several studies (e.g., Greenhalgh and Peacock, 2005; Wohlin, 2014). This step added 17 relevant documents to the dataset.

A flowchart of this process, as well as the number of papers included in each step in the main study is shown in Figure 1.

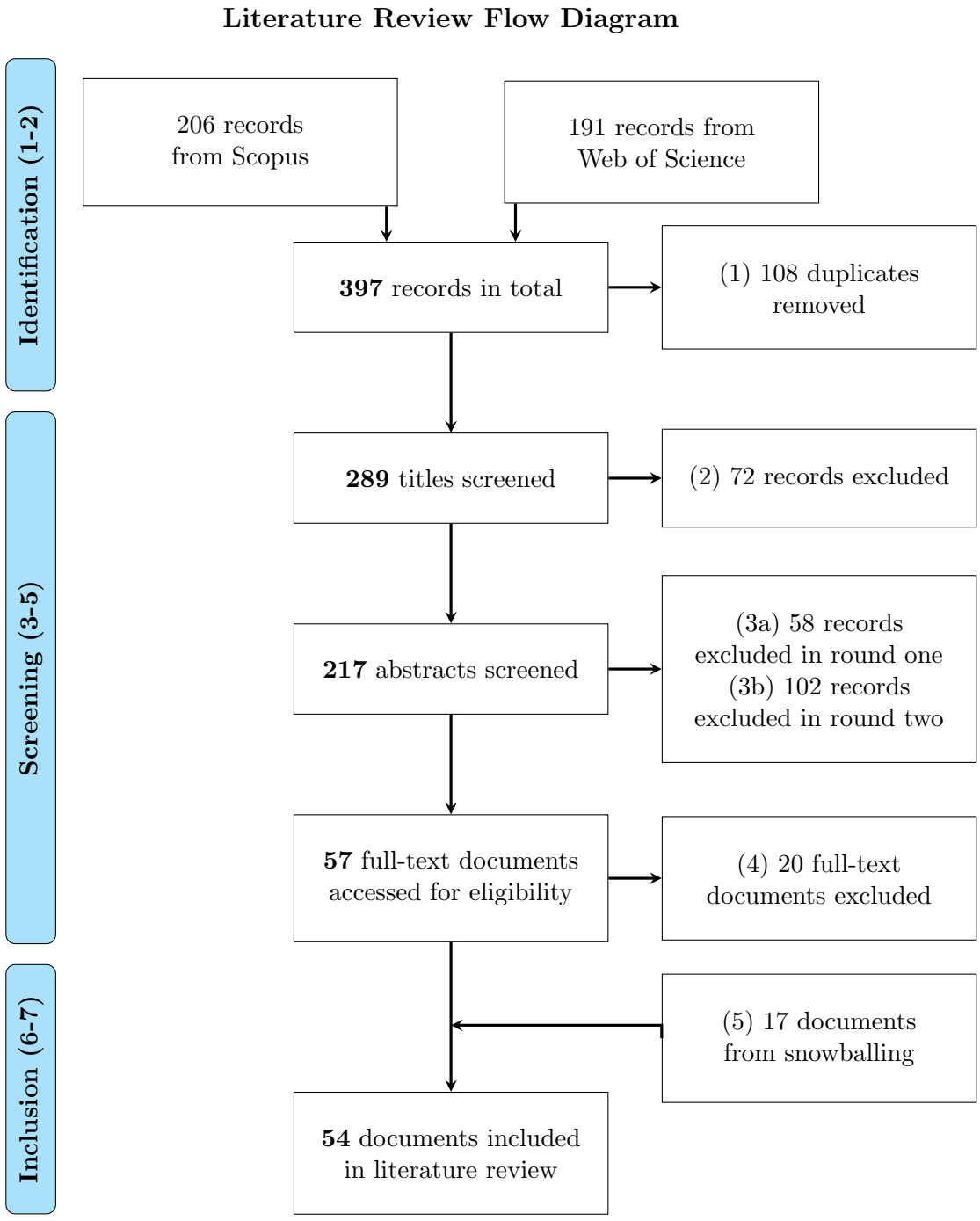

Figure 1: Overview of literature review. Adapted from Page et al. (2021).

The selection of studies followed the eligibility criteria shown in Table 2. One of the criteria – number 10 – was only used for filtering based on titles, in order to include as many relevant papers as possible. The rest of the criteria were used in all steps of the process.

Table 2: Eligibility criteria used for the literature review.

#	Criteria
1	**Include** documents introducing responsible guidelines or general frameworks for responsible AI.
2	**Include** documents focusing on defining responsible AI.
3	**Include** analyses of existing, general frameworks for responsible AI.
4	**Include** documents discussing antecedents for responsibility of AI systems.
5	**Include** documents discussing business advantages of Responsible AI systems.
6	**Exclude** documents of a technical nature, or introducing software systems.
7	**Exclude** documents discussing general AI ethics, or components of responsible AI (such as bias or explainable AI) without introducing frameworks, guidelines or criteria for responsible AI in general.
8	**Exclude** documents focusing on general (non-responsible) AI.
9	**Exclude** documents focused on implementations of AI on specific cases.
10*	**Include** documents discussing ethical AI, ethics in AI or AI governance.

* Only used for screening titles.

Study quality assessment According to Tranfield et al. (2003), quality assessment of management papers can be conducted by either evaluating the methodology of the assessed paper, or by using the rating of the journal it was published in. For this review, it was decided that any journals with a rank of 0 from the Norwegian Register for Scientific Journals, Series and Publishers (Norwegian Register, n.d.) were removed, as the rank indicates that a journal was rejected from the register. This led to one paper being rejected during full-text filtering (step 4).

Due to limitations in the author's language skills, only English papers were to be used in the review. This led to one paper being rejected during full-text filtering (step 4), as it was written in Spanish.

Additionally, it was decided that papers that were unavailable at NTNU, even with help from the libraries, were to be excluded from the review, for natural reasons. This led to one paper being rejected during full-text filtering (step 4).

The rest of the papers rejected during full-text filtering ($n = 17$) were rejected for failing to fulfill the eligibility criteria. 5 of these contained no information, due to

being e.g. workshop- or keynote summaries, while the other 12 were published paper that simply failed to meet the criteria.

In the end, 54 documents were included in the review.

Data extraction and monitoring progress Data was extracted from the documents using an Excel document with pre-defined columns for coding. This method was first tested on a subset of 10 randomly selected documents, which led to the creation of two new columns – *Cases* and *Other relevant data*. The final data extraction form thus consisted of five categories – Metadata, Key concept definitions, Research questions, Added after initial test and Own summary – and a total of 25 columns. An overview of the extraction form, as well as a description of each column, is shown in Table 3. An example of an applied data extraction form is shown in Table 10 (Appendix B).

Table 3: Description of columns used for data extraction

Column	**Description**
Metadata	
Author	Author of the paper
Title	Title of the paper
Source	Journal or publisher of the paper
Source rank	Rank of the source, according to Norwegian Register (n.d.)
Year	Publication year
Abstract	Abstract of the paper
Methodology	Methodology used by the paper
Sample	Sample used, where relevant
Context	Context of the paper
Contribution	Contribution of the paper
Key concept definitions	
AI	Definition of AI
Responsible AI	Definition of responsible AI
AI ethics	Definition of AI ethics
Synonyms	Synonyms for responsible AI used in the text
AI system	Definition of AI system
Research questions	
Principles	Principles for responsible AI used in the paper, with description

Continued on next page...

Column	Description
Antecedents	Antecedents for responsible AI mentioned in the paper
Outcomes	Non-business related outcomes of using responsible AI
Business advantages	Business advantages of responsible AI mentioned in the paper
Barriers	Barriers preventing responsible AI mentioned in the paper
Facilitators	Facilitators for responsible AI mentioned in the paper
Enabling	How to enable principles
Added after initial test	
Cases	Cases of responsible- or irresponsible AI systems
Other relevant data	Catch-all for relevant data that falls outside of these categories
Own summary	
Summary	Own summary of the paper, highlighting relevant information

Data synthesis After extraction, data synthesis was conducted using meta-synthesis (Tranfield et al., 2003) to compare the principles, antecedents and advantages introduced by the different papers, and how they align to solve the same overarching task of making AI responsible. As meta-synthesis aims to "identify theories, [...] or interpretative translations [...] from qualitative studies" (Tranfield et al., 2003, p. 218), this was deemed to be an acceptable method for synthesizing the data.

Due to most of the reviewed papers being of a qualitative nature, it was deemed that meta-analysis or realist synthesis (Tranfield et al., 2003) were inappropriate methods for synthesizing this review. Although meta-ethnography (Tranfield et al., 2003) was considered as a potential methodology for the principles used to answer RQ1, it was quickly found that Ryan and Stahl (2021) provided an exhaustive categorization of principles, thus negating the potential benefits to be gained from the open coding-approach of meta-ethnography.

3.3 Stage III – Reporting and dissemination

The goal of Stage III is to create a comprehensive summary of the findings from the literature review, in order to make it easy for practitioners and other researchers to understand the current state of research within a field (Tranfield et al., 2003).

The report and recommendations The report, i.e., the summary of the findings, should consist of two parts – a descriptive analysis of the field, describing distribution of metadata such as geographical background and age of the research, and a thematic analysis, describing themes and findings from the reviewed papers (Tranfield et al., 2003).

A descriptive analysis of the reviewed papers is included in Section 4.1. This analysis looks at the distribution of publishing year (i.e., age), methodology, geographical context, domain and journal rank of the reviewed papers. A thematic analysis, evaluating the definitions, principles, antecedents and business advantages included in the reviewed papers is included in the rest of Section 4.

Getting evidence into practice Tranfield et al. (2003) argue that an important feature of systematic literature reviews is converting findings into actual implementations used in practice. This paper attempts to answer this by introducing a framework for implementing responsible AI principles in practical development of a project. This framework is described in Section 5.2. This paper also discuss similarities between the reviewed papers, and summarizes methods for implementing the principles found during the review. Evaluating the proposed framework, and the implementation methods, is outside the scope of this paper.

4 Findings

The findings of this section are structured as follows. First, descriptive results are given, describing metadata of the reviewed articles. Then, Section 4.2 contains definitions of key concepts, derived from the reviewed papers. Finally, RQ1, RQ2 and RQ3 are answered in Section 4.3, 4.4 and 4.5, respectively.

4.1 Descriptive Results

As shown in Figure 1, a total of 289 unique titles were assessed during the literature review. After filtering on titles, abstracts, and the full text, and after adding sources from snowballing, 54 documents were included in the review. An overview of the reviewed papers can be seen in Table 11 (Appendix C).

Table 4 shows the yearly distribution of the assessed papers, for each step of the literature review. The oldest of the assessed papers was published in 2015, with

the majority of publications happening in 2019-2022 (n = 279; 97%). The same distribution is reflected in the reviewed papers, where the oldest paper was published in 2017, and 94% of the papers (n = 50) were published between 2019-2022. This indicates a young field of science, that has just recently gained traction. Although the field is still young, discussions with my supervisor led to the conclusion that 289 unique results in the preliminary search means the field has been explored enough to conduct a systematic literature review.

Table 4: Yearly distribution of the assessed papers.

Year	Raw data	After title filtering	After abstract filtering	Review
2015	1	0	0	0
2016	1	0	0	0
2017	0	0	0	1
2018	2	2	1	3
2019	16	14	3	8
2020	45	34	9	10
2021	113	80	24	17
2022	105	83	20	15
2023	1	1	0	0
No year*	5	3	0	0
Total	**289**	**217**	**57**	**54**

* No year provided in the databases.

While the field is relatively young, the methodology used by researchers within the field, as shown in Table 5, indicates a certain level of maturity. Keathley-Herring et al. (2016, p. 930) states that "a research area should evolve from an exploratory beginning to conceptual frameworks being proposed and tested, and then to industry exposure and finally, a convergence on best practices and consistent terminology." The papers included in the review employ a wide range of methodologies, of both an exploratory (reviews, expert discussions), conceptual (conceptual frameworks, viewpoints) and evaluative (qualitative interviews, quantitative surveys, case studies) nature. Some authors (e.g., Fjeld et al., 2020; Jobin et al., 2019) have noted a convergence within the proposed principles as well, indicating that the field is approaching an agreed-upon set of best practices.

The fact that the research being done shows a high level of maturity while the actual field is young may be due to the connections between responsible AI and more established fields within AI- and software ethics. This is further discussed in Section 5.

Table 6 shows the geographical context of the reviewed papers. Over half the papers

Table 5: Methodologies used in the reviewed papers.

Methodology	**Count**	**Percentage***
Review of guidelines	9	17
Literature review	9	17
Qualitative interviews	9	17
Conceptual	7	13
Unknown	6	11
Quantitative survey	5	9
Viewpoint	4	7
Case study	4	7
Expert discussion	3	6
Review of AI strategies	2	4
Review of laws	1	2

* Percentage of reviewed papers that use the methodology. Note that some papers use more than one methodology, so the percentages do not add up to 100.

($n = 32$; 59%) are created without a given context. This is relatively natural, as conceptual papers, viewpoints and literature reviews are rarely limited to geographical locations. All the reviewed papers are written in English, and the included literature reviews are also predominantly reviews of English papers. This creates a possibility for a Western bias, even in papers that appear to be written without a specific context. For the rest of the papers, most are focused on a Western context ($n = 11$; 21%), with some work being done in China ($n = 2$; 4%), India ($n = 2$; 4%) and South Africa ($n = 1$; 2%).

Table 6: Geographical context of the reviewed papers.

Location	**Count**	**Percentage***
Europe	8	15
Global	6	11
The US	2	4
China	2	4
India	2	4
Western countries in general	1	2
South Africa	1	2
N/A†	32	59

* Percentage of reviewed papers that focuses on the given location.
† No context given. Includes viewpoints, conceptual papers, etc.

As with geographical context, most of the reviewed papers ($n = 36$; 67%) are written without a specific domain in mind. This is, once again, natural for conceptual papers, viewpoints and literature reviews, of which there are many among the reviewed papers. The rest of the papers are primarily focused on healthcare ($n = 7$;

13%). Healthcare is a domain with a wide use of technological solutions (e.g. Azaria et al., 2016; Martinez et al., 2008; Son et al., 2014), including AI-based solutions (e.g. Kumar et al., 2023; Singh and Sharma, 2023). As health records are digitized (Rajkomar et al., 2018), yet more avenues for adopting AI opens up, with estimates valuing the data of NHS patients in the UK at around 10 billion GBP per year (Downey, 2019). At the same time, healthcare uses large amounts of personal data, thus increasing the sensitivity patients feel when interacting with systems, automatic or not (Gupta et al., 2021). Professionals in the field of healthcare are therefore used to ethical principles and guidelines, such as the Hippocratic Oath (Wiesing, 2020), and calls have been made to use medical ethics as a starting point for responsible AI guidelines (Dalton-Brown, 2020; Siala & Wang, 2022).

Outside of healthcare, papers tend to look at laws and governance ($n = 3$; 6%) and AI developers ($n = 3$; 6%). Both these cases are natural focus areas for responsible AI. As will be discussed in Section 4.4, the primary antecedent for responsible AI is existing laws and regulations, with national AI strategies also working to push AI systems towards responsibility. Therefore, focusing on how laws and governance impacts – and should impact – the field of responsible AI is important, in order to ensure laws and strategies provide the best possible protection of laypersons, while limiting innovation as little as possible. In a similar fashion, the connection between responsible AI and AI developers is natural. Developers are those who actually design, create and maintain AI systems in use, and are therefore a natural source for understanding current practices in the industry, as well as which principles can be adopted, and how they should be adopted, in order to have the largest impact on moving AI in a responsible direction. A complete overview of the domains used in the reviewed papers can be seen in Table 7.

Table 7: Domains of the reviewed papers.

Domain	**Count**	**Percentage**[*]
Healthcare	7	13
Laws and governance	3	6
Developers	3	6
Labor	1	2
Education	1	2
Finance	1	2
Water	1	2
Multi-sector	1	2
N/A[†]	36	67

[*] Percentage of reviewed papers that focuses on the given domain.
[†] No domain given. Includes viewpoints, conceptual papers, etc.

The reviewed papers come from a wide variety of sources, including journals from several different fields, books, and proceedings. An easy way to quickly assess the quality of a paper is by evaluating the rank of its journal or publisher (Keathley-Herring et al., 2016). In Norway, this is primarily done by the Norwegian Register for Scientific Journals, Series and Publishers, who uses a 3-level ranking system, with level 2 being the highest level, and 0 meaning that a journal was rejected from the register (Norwegian Register, n.d.). Using this register, Table 8 shows the ranks of the reviewed journals. No 0-ranked journals means that the papers can be considered good enough quality to use for a review.

Table 8: Journal ranks of the reviewed papers.

Journal rank	**Count**	**Percentage**[*]
2	6	11
1	34	63
0	0	0
N/A[†]	14	26

[*] Percentage of reviewed papers from a journal with the given rank.
[†] Journal rank not available. Includes books, proceeding papers and journals that have not been evaluated by the Norwegian Register.

Of the 54 papers used for this review, two are from sources that are not peer reviewed – Fjeld et al. (2020) and Floridi and Cowls (2019). Both these sources were gathered during snowballing, and are included for the same reason. Fjeld et al. (2020) compares 36 AI principles, and is considered "a landmark in the synthesis of AI ethics principles and guidance" (Bélisle-Pipon et al., 2022, p. 2). Likewise, Floridi and Cowls (2019) reviewed six sets of principles, and supports the conclusions drawn by Floridi et al. (2018). As the papers are central for the field, and referenced in multiple of the other papers, they were deemed relevant enough to be included in this review.

Some papers may appear to meet the inclusion criteria, but were still included. Examples of this include El-Haddadeh et al. (2021) and Trocin et al. (2021). El-Haddadeh et al. (2021) presents an in-depth comparison of AI used for healthcare in the UK (NHS Test and Trace) and Quatar (EHTERAZ), looking at methods used for achieving responsibility in the two systems, and how well the systems performed at stopping the spread of COVID-19. Although the paper looks at ways AI can be used ethically, it neither defines responsible AI, contains or critizises a framework nor discusses antecedents or business advantages of responsible AI, and therefore does not answer the eligibility criteria. Trocin et al. (2021) performs a literature review aimed at ethical issues in healthcare, and how AI may help solve or mitigate these issues. While the topic of the article revolves around using AI ethically, its

focus is primarily on ethical issues of healthcare, rather than ethical issues of AI, and the paper therefore does not answer the eligibility criteria. The rest of the papers rejected during full-text assessment (step 4, Figure 1) were rejected for similar reasons.

4.2 Key concepts

Although theoretical and philosophical definitions fall slightly outside of the intended scope of this paper, it is necessary for a good understanding of responsible artificial intelligence to look at underlying terms and ensure clearly defined definitions related to the field. While many of the reviewed papers contain interesting and well-thought-out definitions, most of them wary significantly. There is thus a need to unify the definitions found in the field. To do so, this paper use the definitions found during the review, as well as some secondary material, to create classical definitions (Seppälä et al., 2014, p. 36) of the terms *artificial intelligence*, *artificial intelligence system* and *responsible artificial intelligence*.

4.2.1 Artificial intelligence

Not all the reviewed papers define the term artificial intelligence, and the definitions that are used differ in both wording and inclusion. Some similarities exist, however. Common among the definitions given by the reviewed papers is the notion that artificial intelligence technology is based on computing and software, e.g., "...algorithm-driven computing technology..." (Siala & Wang, 2022, p. 1), "...computer systems..." (Brand, 2022, p. 130), "A (computational) technology..." (Dignum, 2021, p. 2), "Autonomous or intelligent software..." (Havrda & Rakova, 2020, p. 1), "...techniques and approaches of computer science..." (Liu et al., 2021, p. 3), and "Machine-based systems..." (Lukkien et al., 2021, p. 1). This largely aligns with the original idea of artificial intelligence, which was based on "a shared vision that computers can be made to perform intelligent tasks" (Moor, 2006, p. 87).

Outside of this, however, the definitions differ. For instance, Mikalef et al. (2022, p. 258) builds on Mikalef and Gupta (2021) and define AI as "the ability of a system to identify, interpret, make inferences, and learn from data to achieve predetermined organisational and societal goals," thus adopting a definition removed from technology but based on learning. The notion of learning is repeated by Dignum (2021, p. 2), who defines AI as "A (computational) technology that is able to infer patterns

and possibly draw conclusions from data." Her definition differs from that of Mikalef et al. in that she focuses on the technology, rather than the system, and she does not include the notion of the system's goals.

Focusing strongly on the outcome and technological side of the technology, Brand (2022, p. 130-131) bases his definition on the Oxford dictionary definition of AI, defining it as "the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages." In a similar vein, Havrda and Rakova (2020, p. 1) adopts the broad definition of AI given by IEEE (2016) as "autonomous or intelligent software when installed into other software and/or hardware systems that are able to exercise independent reasoning, decision-making, intention forming, and motivating skills according to self-defined principles."

Finally, several of the included papers focus on an ability of simulating humans. Liu et al. (2021, p. 3-4) defines AI as "A broad range of techniques and approaches of computer science to simulate human intelligence in machines that are programmed for thinking like humans and mimic human behaviours, capable of performing tasks." Similarly, van Bruxvoort and van Keulen (2021, p. 3) adopt the definition given by Shubhendu and Vijay (2013), and define AI as

> the study of ideas to bring into being machines that respond to stimulation consistent with traditional responses from humans, given the human capacity for contemplation, judgment and intention. Each such machine should engage in critical appraisal and selection of differing opinions within itself.

This paper bases its definition of artificial intelligence on the work done by Siala and Wang (2022, p. 1), who define it as "[A]n algorithm-driven computing technology, [...] programmed to self-learn from data and make intelligent predictions and real-time decisions through the use of artificial neural networks, machine learning, robotic process automation, and data mining." This describes the goals of AI ("to self-learn from data and make intelligent predictions and real-time decisions"), but includes information about how it should be implemented, which limits its use. Removing the implementation details, this paper adapts the work of Siala and Wang (2022) and define *artificial intelligence* as

> An algorithm-driven computing technology, programmed to self-learn from data and make intelligent predictions and real-time decisions.

This definition of artificial intelligence covers several parts of the above-mentioned definitions. The notion of learning from data, as put forth by Mikalef et al. (2022) is explicitly covered. Similarly, the definition used by Dignum (2021) is fully covered, as this paper's definition is based on technology, learns from data and explicitly draws conclusions based on it.

Due to the spread of definitions used by the reviewed papers, adopting a definition from one of them will naturally lead to it differing from others. This definition differs from Liu et al. (2021) and van Bruxvoort and van Keulen (2021) in that it does not require the AI to be able to simulate human intelligence, nor must it have a capacity for contemplation, judgement or intention. Although this paper's definition is less strict than theirs, this brings it more in line with the current reality of AI, as systems able to simulate human intelligence are still more science fiction than reality (Chen, 2020). Similarly, this paper's definition does not mention any specific tasks it should be able to do, which differs from the definitions of Brand (2022) and Havrda and Rakova (2020), but ensures that the paper's definition is applicable to a wider range of systems than theirs.

The rest of this paper will be based on this definition.

4.2.2 Artificial intelligence system

Definitions of AI systems are sparse within the reviewed literature, although some attempts are made. Dignum (2021) bases her definition of an AI system directly on the definition of AI, and defines it as

> A software system (possibly embedded in hardware) designed by humans that, given a complex goal, is able to take a decision based on a process of perception, interpretation and reasoning based on data collected about the environment and that meets the properties of:
>
> - autonomy, meaning that the system is able to deliberate and act with the intent of reaching some task-specific goal without external control
> - adaptability, meaning that the system is able to sense its environment and update its behaviour to changes in the environment
> - interactivity, meaning that the system acts in a physical or digital dimension where people and other systems coexist. (Dignum, 2021, p. 2)

Conversely, Merhi (2022, p. 8) looks at what an AI system cannot do, and defines them as "robots and machines that have a limited ability and lack the intelligence that a human has to make a decision," while van Bruxvoort and van Keulen (2021, p. 2) adopts a "socio-technical perspective" and "view the whole system of the algorithm as a technical component implemented in a technical infrastructure and an organisational structure accompanied by associated rules and regulations."

This paper adopts the definition of AI systems as given by Taddy (2018). He argues that AI systems consist of three pillars that are required for the system to work – a domain structure, data generation and machine learning routines. The domain structure is where the problem an AI system should solve is split into separate parts, that are manageable for simpler machine learning routines. This step requires domain knowledge and business expertise, as the problem must be broken down without inferring with the system's ability to solve it.

The data generation stage is where the AI system receives the data it needs to learn. According to Taddy (2018, p. 4), AI systems "require an active strategy to keep a steady stream of new and useful information flowing into the composite learning algorithms." This stage thus handles all data flow, including collecting it from its own processes.

The final stage, the machine learning routines, take the data generated from the data generation and uses it to learn to solve the broken-down problem provided by the domain structure (Taddy, 2018). This step is therefore what this paper defined as *artificial intelligence*, as it learns from the data it is supplied and makes predictions and decisions.

The difference between AI and AI systems is thus that an AI system incorporates AI in a wider system. Whereas AI by itself passively learn from data and make predictions, AI systems collect data, and use this data with AI to make decisions suitable for the domain they operate in. Where AI is only able to passively learn what is given to them, AI systems are able to operate on their own, making decisions and predictions based on the domain they operate in and the decisions they are given.

To summarize, this paper defines an *artificial intelligence system* as

> A technology system that uses AI to make decisions and predictions that are suitable for the domain the system operates in, given the inputs it is given.

4.2.3 Responsible artificial intelligence

Following the previous two subsections, the only thing that remains in order to understand how to develop responsible AI is a proper definition of the term. The definitions of responsible AI used by the reviewed papers vary slightly. For instance, Doorn (2021, p. 6) defines responsible AI simply as AI that is "consistent with important ethical values," whereas Y. Wang et al. (2020, p. 4964) define it as a human-centered governance framework, used to "harness, deploy, evaluate, and monitor AI machines to create new opportunities for better service provision." In a similar vein, Papagiannidis, Mikalef et al. (2022, p. 58-59) get their definition from Singapore Personal Data Protection Commission (2020), who defines responsible AI as "the process of designing, developing, and deploying artificial intelligence with the purpose of enabling individuals and organisations while also having a fair effect on customers and society."

Several of the reviewed papers use a set of principles as the core of their definition (e.g., Cheng et al., 2021; W. Wang et al., 2021). For example, Barredo Arrieta et al. (2020, p. 83) define responsible AI as "a series of AI principles to be necessarily met when deploying AI in real applications." In a similar fashion, Mikalef et al. (2022, p. 258) builds on Accenture (n.d.), and define responsible AI as "a set of principles that ensure ethical, transparent, and accountable use of AI technologies." Such principles typically make up the core of tools used to achieve responsible AI, and will be further studies later in Section 4.3. However, other tools for achieving the goals of responsible AI exist (Werder et al., 2022), and calls have been made for moving away from principles in the pursuit of responsibility (Henriksen et al., 2021).

Instead, inspiration for a definition can be found in dictionaries. Cambridge Dictionary (*RESPONSIBLE*, n.d.) has, among others, the following definition for the word *responsible*: "having good judgment and the ability to act correctly and make decisions on your own." Using this definition as foundation, it is clear that a responsible AI system must have an idea of what constitutes acting in a "correct" way, and do so to the best of its ability. This leaves the act of concluding what constitutes "correct" acting. In the reviewed papers, this is largely done by a set of principles, which will be further discussed in Section 4.3. The notion of what is a "correct" act, however, can vary between different technical systems (Hagendorff, 2020), as well as different cultures and background (Ford & Richardson, 1994). Although some attempts have been made (e.g. Hagendorff, 2020; Jobin et al., 2019), no universally agreed-upon set of principles exist.

In order to create a globally acceptable definition of responsible AI, this paper instead bases its definition of responsible AI on two views of applied ethics – a utilitarian view, which simplifies a "correct" act to one that leads to maximum global utility (Lang, 2004), i.e., the largest possible amount of "good" for the largest possible amount of people, and a negative utilitarian view, which argues that a "correct" act is the one that leads to the least global suffering (Gandjour & Lauterbach, 2003). Combining these two views lead to defining *responsible artificial intelligence* as

> An AI system designed, developed and used in an ethical way, to maximize global utility and minimize global harm.

As will be further discussed in Section 4.3, maximizing utility and minimizing harm are two similar, but not identical, goals. The goal of both is the same, however, in that "correct" actions are those that focus on "the greater good"

This definition is largely in line with the definition used by Doorn (2021), as both utilitarianism and negative utilitarianism are established ethical principles, and AI systems that follows these thus act consistent with ethical values. Similarly, if a service provides utility, then creating "new opportunities for better service provision" can be considered a way of increasing global utility, thus aligning this definition with that used by Y. Wang et al. (2020). The same, naturally, holds if the original system reduces harm. By acting fair, an AI system can reduce harm in the form of discrimination, and by enabling individuals, the system increases utility. This definition is, thus, also in line with that used by Papagiannidis, Mikalef et al. (2022).

The definition differs from all those based on principles, as it also approves of AI systems acting responsibly based on other methods. However, ethical principles still make up the core of current methods for developing responsible systems. This paper therefore considers ethical principles as criteria for responsibility, and use them as the core of the implementation of responsibility, even for responsible AI systems following this definition.

It is important to note that throughout this paper, responsible AI will be used in contrast to regular AI system. As such, any antecedents, as discussed in Section 4.4, are antecedents for adopting responsible principles, i.e., converting existing AI systems to be responsible, rather than antecedents for implementing AI solutions in the first place. The same is true for business advantages discussed in Section 4.5.

4.3 RQ1 – Principles for responsible AI

Of the 54 reviewed papers, 48 (89%) present or use a set of principles to create or describe responsible AI, for a total of 56 unique principles. Three steps are used to make this set of principles more manageable. First, principles were clustered following the categorization done by Ryan and Stahl (2021). Their work reviewed existing guidelines, and came up with 11 categories – Transparency, Justice and fairness, Non-maleficence, Responsibility, Privacy, Beneficence, Freedom and autonomy, Trust, Sustainability, Dignity, and Solidarity – containing a total of 61 subcategories, which most of the reviewed principles aligned with. The principles that did not align with any subcategory were categorized with the closest fit, with principles that did not align with any clusters being grouped as *Others*. This ensured all principles would be used in the review, without creating multiple low-density clusters.

It is important to note that there may be some overlap between the principles. An example of this can be seen in Fjeld et al. (2020), where the principle of *Accountability* includes as Environmental Responsibility as a subcategory, or in Mikalef et al. (2022), where the principle of *Societal and environmental well-being* is included in both *Beneficence* and *Sustainability*. In these cases, principles are either included in both clusters, which is the case with Mikalef et al. (2022), or in the one that aligns the most, as done with Fjeld et al. (2020), depending on how much of the principle overlaps.

After clustering the principles, some changes were made to the categories proposed by Ryan and Stahl (2021). Clusters with fewer than five mentions among the reviewed texts were removed, and their principles moved to other clusters. This meant that *Dignity* ($n = 3$) was moved to be part of *Non-maleficence*, and *Solidarity* ($n = 1$) became part of *Justice and fairness*. As dignity can be considered to be a harm, and thus can be avoided by ensuring AI systems can do no harm, and as solidarity can be created by ensuring systems act in a fair and just manner, this change felt like a natural decision. Then, some categories were renamed – *Justice and fairness*, and *Freedom and autonomy* were simplified to *Justice* and *Autonomy*, respectively. This was done to better aligned with common usage throughout the reviewed papers (e.g., Balagué, 2021; Floridi et al., 2018, and the papers basing their principles on these).

After clusters had been created, these were again divided into two sets – *core* and *instrumental* principles. This division was argued by Canca (2020), in order to separate core principles, that are "intrinsically valuable" (p. 19), from instrumental principles, "whose values are derived from their instrumental effect in protecting and

promoting intrinsic values" (p. 20). The motivation for this was to align principles of AI ethics with a "widely utilized set of core principles in applied ethics" (Canca, 2020, p. 19).

One key change was done to the division suggested by Canca (2020). *Beneficence*, as suggested by Canca (2020) to encapsulate both "avoiding harm" and "doing good" (p. 19), was split into *Beneficence* (doing good) and *Non-maleficence* (avoiding harm). This has three benefits. First, it brings the division in line with the categories proposed by Ryan and Stahl (2021), which are already used for the clustering mentioned above. Secondly, this aligns core principles of responsible AI with those used in other fields of applied ethics, such as bioethics (Beauchamp & Childress, 2001) and medicine (Gillon, 1994, 2003). Finally, this resonates with principles used by several of the reviewed papers (e.g., Balagué, 2021; Floridi et al., 2018; Jobin et al., 2019; Nauck, 2019), ensuring principles that are actually used within the field of responsible AI are placed in the group they belong.

The distribution of principles used in the reviewed papers is shown in Figure 2.

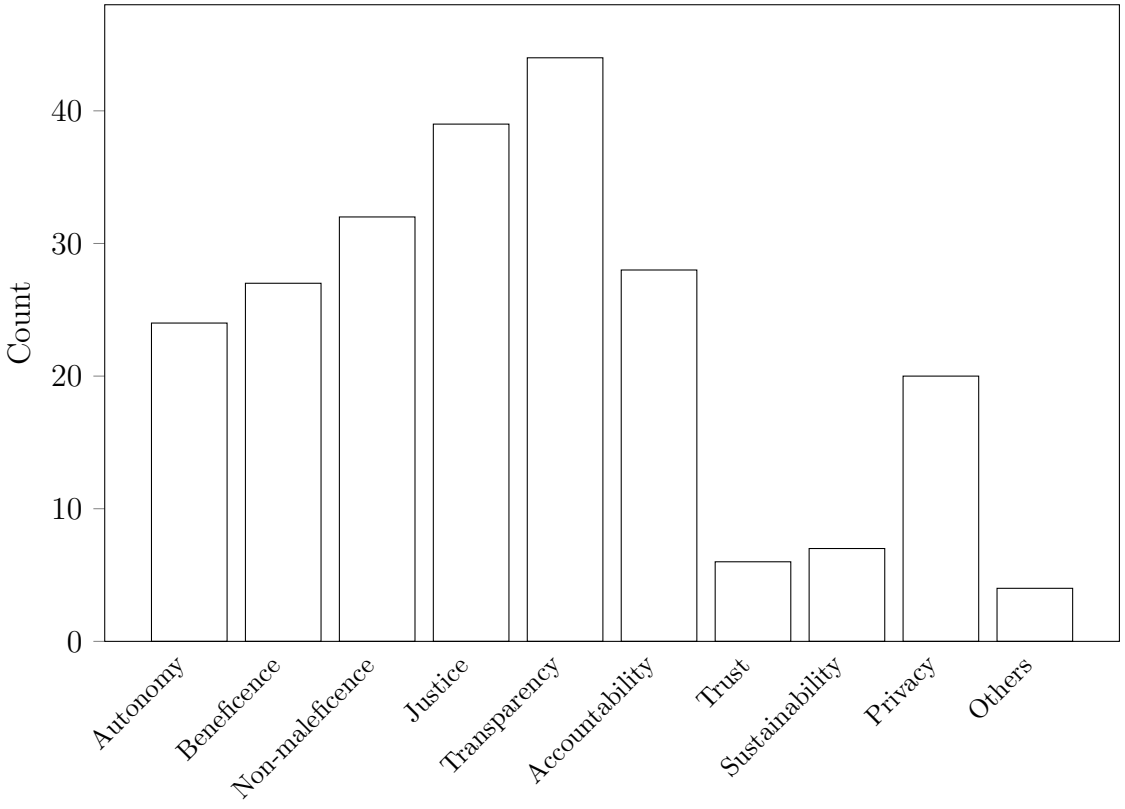

Figure 2: Distribution of principles in the reviewed papers

4.3.1 Core principles

Core principles are described by Canca (2020) as principles that "invoke those values that theories in moral and political philosophy argue to be intrinsically valuable, meaning their value is not derived from something else" (p. 19). She argues that core principles, then, should be present in every responsible AI system, and that they – by themselves – should be used to select the correct action for a given context, as the action that fulfills the core principles to the highest degree (Canca, 2020).

As mentioned above, this paper adapts the core principles suggested by Canca (2020) to be the following clusters of principles: *Autonomy*, *Beneficence*, *Non-maleficence* and *Justice*. A complete overview of the core principles used by the reviewed papers is shown in Table 12 (Appendix D).

The notion that core principles should be present in every responsible AI system is somewhat reflected in the reviewed papers. Of the 48 papers that include a set of principles, 18 of them (38%) include all four core principles. In comparison, only 6 (13%) contain none of the core principles. Perhaps most importantly, most sets of principles created by reviewing existing guidelines for responsible AI contain all four (Clarke, 2019; Floridi et al., 2018; Jobin et al., 2019; Ryan and Stahl, 2021) or three (Fjeld et al., 2020) core principles. The only exceptions from this are Brand (2022) and Hagendorff (2020), who only include one core principle each. Brand (2022) is focused on developing a framework for responsible AI policy and regulation, and this may be explained by him primarily targeting more easily measureable principles, such as transparency and privacy. Hagendorff (2020), however, presents an interesting case. Although he only includes one of the core principles (*Fairness*, which is included in *Justice*) in his list of recurring principles, his overview (Hagendorff, 2020, Table 1) shows that both "common good" (*Beneficence*), "human oversight" (*Autonomy*) and "safety, cybersecurity" (*Non-maleficence*) are commonly used among the guidelines he reviews. This indicates that core principles of responsible AI are extensively used in existing guidelines in the AI industry.

Autonomy Autonomy can be defined as "the quality or state of being self-governing" and "self-directing freedom and especially moral independence" (*Autonomy*, n.d.). In relation to responsible AI, autonomy can be thought of as "the idea that individuals have a right to make decisions for themselves about the treatment they do or not receive" (Floridi et al., 2018, p. 697-698). Following the categorization done by Ryan and Stahl (2021, Table 1), *Autonomy* includes the subcategories Freedom, Autonomy, Consent, Choice, Self-determination, Liberty, and Empowerment.

Autonomy is included in 24 papers (50% of all principle-including papers).

Rothenberger et al. (2019, p. 5) argue that "An AI should have a purpose," and that this purpose should be to support, and not replace, humans. Thelisson (2018, p. 387), Q. Lu et al. (2022, p. 102). Fjeld et al. (2020, p. 53) and Clarke (2019, p. 416) all stress that AI should stay under human control. Among the guidelines reviewed by Jobin et al. (2019, p. 11), autonomy is referred to as both a "positive freedom," i.e., a freedom to do good things, which includes "the freedom to flourish, to self-determination through democratic means, the right to establish and develop relationships with other human beings, the freedom to withdraw consent, or the freedom to use a preferred platform or technology," as well as a "negative freedom", i.e., a freedom from bad things, such as "freedom from technological experimentation, manipulation or surveillance." According to the guidelines reviewed by Floridi et al. (2018, p. 698), a key part of ensuring autonomy of humans is by limiting the autonomy of machines, ensuring that humans "retain the power to decide which decisions to take."

The reviewed papers contain a wide set of methods to ensure human autonomy in responsible AI systems. First, responsible AI systems should be built to ensure the freedoms listed by Jobin et al. (2019) are not impacted. This means responsible AI systems should not be used for manipulation or surveillance, nor should it limit the freedoms of humans. Secondly, responsible AI systems should ensure people affected by their decisions have a way to "challenge the use or output of the system" (Q. Lu et al., 2022, p. 102), "request and receive human review of those decisions" (Fjeld et al., 2020, p. 53) and "to opt out of automated decision" (Fjeld et al., 2020, p. 54). AI developers should ensure informed consent is given before data is collected from people (Jobin et al., 2019; Lukkien et al., 2021), and implement methods for humans to take control and override the automatic system when necessary (Floridi et al., 2018; Rakova et al., 2021). Liu et al. (2021) argue that AI systems should adapt to each user, ensuring each person gets the level of autonomy and assistance as they need.

Some of the reviewed papers (e.g., Brand, 2022; Nevanperä et al., 2021) argue that responsible AI systems should have a "human in the loop." This entails having a human involved in and responsible for the decisions made by the AI system, and comes with two benefits. First, a clear agency for the system is created, as the person is responsible even if decisions are automated (Nevanperä et al., 2021). Secondly, and perhaps more important, this lets a human "perform checks and/or investigate situations with considerable uncertainty" (van Bruxvoort & van Keulen, 2021, p. 12). Manually handling unclear decisions means that the AI stays under

human control, and limits autonomy no more than a fully manual system would.

Beneficence Beneficence is defined as "the quality or state of doing or producing good" (*Beneficence*, n.d.). This means that AI systems fulfilling the principle of beneficence should work towards the greater good, i.e., solutions that are good for humanity. According to Ryan and Stahl (2021), *Beneficence* include the subcategories Benefits, Beneficence, Well-being, Peace, Social good, and Common good. Beneficence is included in 27 (56%) of the papers.

According to van Bruxvoort and van Keulen (2021, p. 9), AI systems fulfilling the principle of *Beneficence* should have the "well-being of humans, society and planet put upfront." Beneficent AI systems are systems created to be beneficial for humanity (Floridi et al., 2018), promoting good (Jobin et al., 2019) and increase societal well-being (Mikalef et al., 2022). Barredo Arrieta et al. (2020, p. 103) argue that responsible AI systems should "be aligned with the United Nation's Sustainable Development Goals and contribute to them in a positive and tangible way." In a similar fashion, Buhmann and Fieseler (2021, p. 2) highlight responsible AI's "responsibility to do good," i.e., it should work towards "improvement of living conditions, such as in accordance with the sustainable development goals." Responsible AI should "complement humans" (Clarke, 2019, p. 419), and "be developed and used to increase prosperity for all," thus advancing "inclusive growth and sustainable development" (Rizinski et al., 2022, p. 97534).

Several methods exist to ensure AI systems are beneficial. On an organisational level, Morley et al. (2020, p. 2151) argue that AI systems should be justified, i.e., "the purpose for building the system must be clear and linked to a clear benefit." To support this, organisations "should define as clear as possible what it views as beneficial to humans [...], so that its view can guide decisions for all aspects" (van Bruxvoort & van Keulen, 2021, p. 9). In other words, AI organisations should create an organisational definition of beneficence, which can be used in their AI development in the same way corporate visions and values can be used. Beneficence should be considered when deciding what problems to solve, and beneficial AI systems should work towards "solutions to some of the world's greatest problems, such as curing diseases, ensuring food security and preventing environmental damage" (Ryan & Stahl, 2021, p. 73). Responsible AI organisations should work towards "minimizing power concentration," "working more closely with 'affected' people," and "minimizing conflicts of interests" (Jobin et al., 2019, p. 11).

Beneficial AI systems should include stakeholders in their decision making (Havrda & Rakova, 2020; Q. Lu et al., 2022; Morley et al., 2020), ensuring that everyone who

are impacted by the system have a say. Nevanperä et al. (2021, p. 134) argue that, due to the wide reach of AI systems, "all sentient beings ought to be considered as stakeholders." In order to properly ensure systems are beneficial, they should include metrics that can be used as key performance indicators to evaluate the beneficence of the system (Eitel-Porter, 2021). These may, for example, be "metrics related to satisfaction with life, affect, psychological well-being, and inequality" (Havrda & Rakova, 2020, p. 5).

Non-maleficence Maleficence is defined as "the act of committing harm or evil," leading non-maleficence to mean the avoidance of committing harm. Floridi et al. (2018, p. 697) define non-maleficence simply as "do no harm." Following Ryan and Stahl (2021), *Non-maleficence* includes the subcategories Non-maleficence, Security, Safety, Harm, Protection, Precaution, Prevention, Integrity, and Non-subversion. Non-maleficence is the second most used core principle amongst the reviewed papers, with 32 papers (67%) including the principle.

Doorn (2021, p. 6) define non-maleficence as "an intention to avoid needless harm or injury," and states that "it is often interpreted as the need to protect safety and security." People creating and using non-maleficent AI systems should assess, control for and avoid or mitigate potentially harmful impacts (Buhmann & Fieseler, 2021; Fjeld et al., 2020), in both the short- and long term (Clarke, 2019, p. 416), and of both an accidental and deliberate nature (Floridi et al., 2018, p. 697). This includes both direct impacts, such as autonomous weapon systems being used for killing (Eitel-Porter, 2021), but also indirect impacts, such as job displacement (Gupta et al., 2021, p. 12) and power concentration (Jobin et al., 2019).

Non-maleficent AI systems should have "resistance to malfunctions (robustness) and recoverability when malfunctions occur (resilience)" (Clarke, 2019, p. 416), resist both internal and external manipulations (Fjeld et al., 2020; Rothenberger et al., 2019), "not diminish or destroy but respect, preserve or even increase human dignity" (Jobin et al., 2019, p. 13), and "ensure data security and AI safety" (Buhmann & Fieseler, 2021, p. 2), through means like data governance (Werder et al., 2022). Anagnostou et al. (2022, p. 13) argue that developers of non-maleficent AI systems should focus "not only security but also cybersecurity," i.e., protection from attacks (Ryan & Stahl, 2021, p. 70).

Non-maleficent AI can be achieved by involving stakeholders (Dignum, 2021; Jobin et al., 2019), using their perspectives in an impact assessment and including them in the design process (Clarke, 2019, p. 416), and to help better identify potential impacts (Havrda & Rakova, 2020). Such impact assessments should be documented

and regularly reviewed (Ryan & Stahl, 2021, p. 70), and create a basis for risk management methods, designed to minimize the risk of harmful consequences (Brand, 2022; Buhmann & Fieseler, 2021; Clarke, 2019). AI systems should be regularly tested, audited and validated (Fjeld et al., 2020; Jobin et al., 2019; van Bruxvoort & van Keulen, 2021), for example by automatic checks of safety metrics developed for the system (Havrda & Rakova, 2020). AI organisations should ensure their data is protected, regularly updated and only available to workers who needs access (Mikalef et al., 2022; van Bruxvoort & van Keulen, 2021). Finally, if the potential harm of a system is too large, organisations should "consider alternative, less harmful ways of achieving the same objectives" (Clarke, 2019, p. 419), such as by using less sensitive data (van Bruxvoort & van Keulen, 2021, p. 10).

Justice Justice can be defined as "the quality of being just, impartial, or fair" (*Justice*, n.d.). In the words of Floridi and Cowls (2019, p. 7), justice in an AI context means "eliminating unfair discrimination, promoting diversity, and preventing the rise of new threats to justice." Once again following the categorization done by Ryan and Stahl (2021), *Justice* includes the subcategories Justice, Fairness, Consistency, Inclusion, Equality, Equity, Non-bias, Non-discrimination, Diversity, Plurality, Accessibility, Reversibility, Remedy, Redress, Challenge, and Access and distribution. Justice is used by 39 papers (81%), and is thus the most used core principle among the reviewed papers.

Just AI systems should minimize the bias found in their data (Clarke, 2019; Fjeld et al., 2020; Jobin et al., 2019; Nauck, 2019; Rothenberger et al., 2019), be inclusive (Q. Lu et al., 2022; Mikalef et al., 2022), encourage diversity (Mikalef et al., 2022) and equality (Canca, 2020; Fjeld et al., 2020; Gupta et al., 2021), and avoid discriminatory results against individuals or groups of individuals (Hacker & Passoth, 2022; Jobin et al., 2019; Q. Lu et al., 2022; Mikalef et al., 2022; Werder et al., 2022). Any negative consequences that cannot be avoided through non-maleficence should be distributed fairly, as should the benefits gained through beneficence (Doorn, 2021; Jobin et al., 2019). AI developers should not only avoid previous bias, such as bias found in existing datasets, but also be aware of the risk of discriminating in new ways and on previously unforeseen scales (Fjeld et al., 2020, p. 48).

To ensure AI systems are just, calls are once again made for involving stakeholders (e.g., Fjeld et al., 2020; Q. Lu et al., 2022), ensuring "voices and data of minority populations" are included in the AI system (Lukkien et al., 2021, p. 10), and that potentially missed biases are mitigated (Siala & Wang, 2022). This can be strengthened by having diverse AI development teams, ensuring a wide range of

genders, ages, races and cultural backgrounds (Q. Lu et al., 2022, p. 104), as well as professional backgrounds and skill sets (Fjeld et al., 2020, p. 52). Just AI systems should ensure they are built on diverse (Doorn, 2021) and representative (Fjeld et al., 2020; Hacker & Passoth, 2022) data, and access to AI systems should be distributed equally and fairly (Canca, 2020; Fjeld et al., 2020; Jobin et al., 2019). Individuals and groups from worse-off backgrounds should be especially protected from harm (Canca, 2020).

Fairness can be controlled with risk identification methods and checklists specifically aimed at fairness (Q. Lu et al., 2022, p. 103). According to Q. Lu et al. (2022), there already exist technical tools, such as "IBM's AI Fairness 360, Google's Fairness Indicators, Microsoft's Fairlearn, and UChicago's Aequitas" (p. 107), and statistical methods, such as "demographic parity, data augmentation, weighted data sampling, re-sampling, re-weighting, swapping labels, removing dependencies, equalized odds checking" (p. 109) for assessing the fairness of a system. Organisations can then show the fairness of their systems by disclosing "summary statistics showing the distribution of scores between different protected groups" (Hacker & Passoth, 2022, p. 365).

4.3.2 Instrumental principles

Whereas core principles are intrinsically valuable, instrumental principles provide ethical value by supporting and facilitating for the core principles (Canca, 2020). Which instrumental principles to include when developing a responsible AI system is therefore up to the developers and designers of the system. As ethics is not absolute, the selection of instrumental principles may vary depending on, amongst others, the cultural context of the AI system (van Bruxvoort & van Keulen, 2021), the support system of the country the AI system is to be used in (Wright & Schultz, 2018), or the ethnicity, gender, political leaning and background of afflicted stakeholders (Jakesch et al., 2022), and may even change over time (van Bruxvoort & van Keulen, 2021).

Based on the clustering described above, the reviewed papers include the following categories of instrumental principles: *Transparency*, *Accountability*, *Trust*, *Sustainability*, *Privacy* and *Others*, where the latter encapsulates Laws and regulations, Predictability, Quality benchmarking, and Interventions and co-design. A complete overview of the instrumental principles used by the reviewed papers is shown in Table 13 (Appendix D).

While this list encompasses instrumental principles used by the reviewed papers, it is by no means an exhaustive list of potential principles designers of AI systems can

use, nor is that the intention of this section. Instead, this section is meant to give an overview of some principles that are used by current literature on responsible AI, in order to show some methods to facilitate for the core principles – the principles that actually make an AI system responsible – mentioned above. While the principles mentioned here are relatively common, and can thus work as a starting point for selecting relevant instrumental principles for a new system, designers should ensure they are adapted to the system and context they operate in, and that relevant principles that are not mentioned here are also included in their projects.

Transparency Transparency is defined by Dignum (2017, p. 2) as the ability to "[understand] the ways systems make decisions and how the data is being used, collected and governed." Transparency is the most used principle among the reviewed principles, being used by 44 papers (92%). According to Ryan and Stahl (2021), *Transparency* includes the subcategories Transparency, Explainability, Explicability, Understandability, Interpretability, Communication, Disclosure, and Showing.

Transparent AI systems should be able to explain their decisions in a way that is understandable for humans (Clarke, 2019), and adapted to the audience it is presented for (Barredo Arrieta et al., 2020; Hacker & Passoth, 2022). It should be clear for users of an AI system that they are interacting with an artificial intelligence, and not a human being (Clarke, 2019; Nauck, 2019), and AI systems should be open about how and why data is collected and used (Clarke, 2019; Jobin et al., 2019). Not only should AI systems be transparent, but AI organisations should be open about the decisions taken during the development of the AI systems (Vakkuri et al., 2022), which protocols were used for development (van Bruxvoort & van Keulen, 2021), and what AI governance methods used to ensure that their system stays responsible (Dignum, 2017, 2019). Transparent AI systems should involve stakeholders in the decisions taken during the design of the system (Dignum, 2019). AI systems could design and incorporate metrics for stakeholder understanding of the system (Havrda & Rakova, 2020), and track these similar to other key performance indicators. Some of the guidelines reviewed by Fjeld et al. (2020) and Jobin et al. (2019) argue that responsible AI systems should open-source their code and data, to fully facilitate public scrutiny.

Modern AI methods, such as deep neural networks and reinforcement learning systems, are complicated mathematical systems, whose decisions are typically hard to explain (Barredo Arrieta et al., 2020). Models based on these methods are therefore often called black box models (Bélisle-Pipon et al., 2022). Several technical tools exist to make such black box models more transparent (see, e.g., Barredo Arrieta

et al., 2020), although these may lead to a trade-off, where increased transparency leads to worse performance (Barredo Arrieta et al., 2020). Calls have therefore been made to abandon well-performing but hard-to-explain black box models, and instead adopt more interpretable models (Barredo Arrieta et al., 2020; Rizinski et al., 2022; Rudin, 2019).

Transparency creates ethical value in two ways. First, it creates a way to audit the core principles (Canca, 2020). This enables both internal and external scrutiny of the system, which can decrease the time that passes before unforeseen harm, injustice or discrimination is detected. Similarly, this can ensure loss of autonomy is quickly noticed, thus decreasing the risk of it becoming a problem. Secondly, transparent systems allow developers to understand how their system works, and thus how it can be improved to provide better results (Barredo Arrieta et al., 2020). This may help offset the trade-off mentioned earlier, thus increase the beneficence of the system.

Not all papers agree with transparency being an instrumental principle. Floridi et al. (2018), and papers using their principles, include Explicability as part of their core principles, arguing that the other principles cannot be put into practice, and especially not evaluated, without having an explicable system. In a similar fashion, Jobin et al. (2019) discuss a claim by Turilli and Floridi (2009), that "transparency is not an ethical principle in itself but a proethical condition for enabling or impairing other ethical practices or principles" (Turilli & Floridi, 2009, p. 105). As *Transparency* can be enabled and used the same as other principles for responsible AI, it is still included in this paper, but this, combined with the broad usage of the principle among the reviewed papers, indicate that transparency should be strongly considered for all responsible AI systems, no matter the context they operate in.

Accountability Accountability is described as "the ability to determine whether a decision was made in accordance with procedural and substantive standards and to hold someone responsible if those standards are not met" (Anagnostou et al., 2022, p. 36). 28 of the reviewed papers (58%) include accountability as part of their principles, making it the second most used instrumental principle. According to the categories of Ryan and Stahl (2021), *Accountability* (which they describe as *responsibility*) includes subcategories Responsibility, Accountability, Liability, and Acting with integrity.

Accountable AI systems should ensure a legal person, or group of legal persons, are held responsible for impacts of the automatic system (Vakkuri et al., 2022; Werder et al., 2022), and for the different parts of the system (Q. Lu et al., 2022).

Such responsibility may be shared between developers and operators of the system (Vakkuri et al., 2022), and can be ensured by including a human in the loop, as discussed as a solution for *Autonomy* (Doorn, 2021). In cases where unintended impacts happen, accountable AI systems should have ways to appeal the automatic decisions, and remedies for issues caused (Clarke, 2019; Eitel-Porter, 2021; Fjeld et al., 2020). In order to enable post hoc analyses of such incidents, decisions made during the development process should be documented and saved, to enable audits of the system (Barredo Arrieta et al., 2020; Havrda & Rakova, 2020; Rizinski et al., 2022). Such audits should be used to improve the system over time (Fjeld et al., 2020).

It is important that accountable AI systems have justifiable (Kumar et al., 2021) and replicable (Fjeld et al., 2020) outcomes, "based on original data" (Merhi, 2022, p. 3) and "derivable from [...] the learning algorithms used" (Dignum, 2017, p. 5). Accountability can be established through certification standards, which are popular among AI developers (Henriksen et al., 2021), or the use of contracts, especially useful for dividing responsibility between developers and customers of a system (Jobin et al., 2019; Vakkuri et al., 2022). In order to be accountable, AI organisations should protect whistleblowers (Jobin et al., 2019), and have a way for workers to veto an AI system if there are any concerns (Anagnostou et al., 2022).

There has been some discussion regarding whether AI systems themselves can be held responsible for their actions, the same way legal persons are (e.g., papers reviewed by Jobin et al., 2019). The consensus, however, is that such responsibilities cannot be placed on the technology, but should instead be placed on the people designing, developing and operating the system (Anagnostou et al., 2022; Doorn, 2021; Fjeld et al., 2020).

Accountability is made possible by transparency, as decisions made during development of the system can be used to assign responsibility (Vakkuri et al., 2022). By enabling audits, accountability minimizes the risk of harm or discrimination going unnoticed, this supporting the core principles of *Non-maleficence* and *Justice*. Accountability further supports the notion of having a human in the loop, or otherwise responsible for decisions made automatically, thus supporting the principle of *Autonomy*.

Trust Public trust in AI systems is essential if they are to reach their full potential (Jobin et al., 2019). Trust was included in 6 of the reviewed papers (13%), including one (Jakesch et al., 2022) who dropped it as a result of their pilot study. This leaves it as the least popular principle among the reviewed papers. According to Ryan and

Stahl (2021), *Trust* only has one subcategory: Trustworthiness.

AI organisations should create trustworthy and reliable systems, and prove this to their stakeholders (Ryan & Stahl, 2021). To create trust, systems should use high-quality data to work as intended and expected (Ryan & Stahl, 2021), and to minimize bias (Y. Wang et al., 2020). Organisations should facilitate for trust "among scientists and engineers," in order to advance the state of AI research (Jobin et al., 2019).

Trust creates value by facilitating for accountability, as the two principles align closely – increasing accountability is likely to increase trust, and increasing trust is likely to enable more feedback on systems, thus maximizing the potential for beneficence, while enabling early warnings of injustice and maleficence.

Sustainability Sustainability in an AI context means that the environmental impact of AI systems should be minimized (Siala & Wang, 2022). Sustainability was used in 7 of the reviewed papers (15%). Ryan and Stahl (2021) give *Sustainability* the subcategories Sustainability, Environment (nature), Energy, and Resources (energy).

Sustainable AI systems should avoid harming the environment more than necessary (Mikalef et al., 2022), and to "increase prosperity for [...] the planet" (Rizinski et al., 2022, p. 97534). As AI models grow in size and depth, they should be "developed in an environmentally conscious manner" (Ryan & Stahl, 2021, p. 75), minimizing their energy consumption and environmental footprint (Jobin et al., 2019; Ryan & Stahl, 2021), such as by using "special hardware designed for energy-efficient learning" (van Bruxvoort & van Keulen, 2021). "AI should not be used to harm biodiversity" (Ryan & Stahl, 2021, p. 75), making it reasonable to consider the environment and animals (i.e., all "living beings" (Havrda & Rakova, 2020, p. 4)) as stakeholders of the system.

Sustainability is largely aligned with non-maleficence and beneficence (Peters et al., 2020), in that positive outcomes of sustainability – such as reduced harm to the environment – leads to AI systems that do good and minimize harm. As "climate change [...] increasingly affect the poor" (Abeygunawardena et al., 2002, p. IX), limiting this can also be seen as a contribution toward global justice.

Privacy The principle of Privacy in an AI context "relates to the use of data and the need to protect people's right to privacy" (Doorn, 2021, p. 6), i.e., people's right to "freedom from unauthorized intrusion" (*Privacy*, n.d.). Privacy was used in 20 of

the reviewed papers (42%). According to Ryan and Stahl (2021), *Privacy* includes the subcategories Privacy, and Personal or private information.

Privacy in AI systems "includes both information provided by users and information generated about those users derived from their interactions with the system" (Barredo Arrieta et al., 2020, p. 106), and both direct (i.e., through access to the data) and indirect (i.e., through access to the model) information exposure (Cheng et al., 2021, p. 1155). Privacy-compliant AI systems should collect as little data as possible while still fulfilling their goal (Doorn, 2021; Havrda & Rakova, 2020; Ryan & Stahl, 2021), and should only do so with prior consent from their data sources (Fjeld et al., 2020; Rothenberger et al., 2019). Data should be stored safely (Liu et al., 2021), and personal data should be removable if the source of the data so wishes (Fjeld et al., 2020; Ryan & Stahl, 2021). AI organisations complying with the principle of privacy should conduct privacy-focused impact assessments (Havrda & Rakova, 2020), and limit access to their data (Jobin et al., 2019). Finally, AI developers following the principle of privacy can implement a range of privacy-preserving technical tools, such as federated learning, decentralized learning (Q. Lu et al., 2022) or de-identification (Ryan & Stahl, 2021), to increase the privacy of their system.

As the harm to avoid in Non-maleficence includes "violation of privacy" (Jobin et al., 2019, p. 9), it is clear that increasing privacy directly leads to value through Non-maleficence. The same connection can be found with Autonomy, where increasing privacy leads to fewer connections between a person and their digital actions or expression, thus allowing freer speech and more individual freedom. As personal data such as race and gender are removed from AI systems, increasing privacy may also lead to less discrimination, thus creating value through increasing Justice.

Others Some reviewed papers highlight principles that do not fit within the definitions of the other clusters. As these are not used widely enough to be considered their own categories, they are instead described separately.

Laws and regulations is included by Merhi (2022) and Mikalef et al. (2022) who argue that "AI systems should adhere to the respective laws and regulations that dictate their functioning" (Mikalef et al., 2022, p. 259). This is especially important when AI is used to automate jobs or disrupting existing markets, where laws and regulations ensure the harm caused is as small as possible (Mikalef et al., 2022).

Predictability is used by Vakkuri et al. (2022) to explain systems that "act in a predictable manner in any given situation" (p. 102). Increasing the predictability of

an AI system can make it easier to understand how it works, and thus creates value through increasing the Transparency of the system – with the increased value that brings (Vakkuri et al., 2022).

Quality benchmarking, as described by Hacker and Passoth (2022), has two benefits. First, disclosing performance metrics of AI systems would give potential customers ways to objectively compare different systems, thus creating competition in the AI market. Secondly, having quality benchmarks allows for better assessment of risk and rewards of AI systems, compared to non-AI systems. Both these benefits show that quality benchmarking creates value by increasing the Transparency of the system it is implemented in.

Finally, *Interventions and co-design*, also used by Hacker and Passoth (2022), includes taking existing methods for creating responsible systems, from fields such as human computer interaction and RRI, to design systems alongside stakeholders and affected persons. Doing so has the possibility of creating value through all core principles – Autonomy and Non-maleficence can be supported by ensuring previously hidden harms and negative outcomes are detected early, Beneficence can be supported by ensuring alignment of stakeholder- and system values and priorities, and Justice can be supported by including minorities and specially affected persons among the involved stakeholders, thus designing a system that does not negatively impact those groups.

4.3.3 Methods for enabling responsible AI

While the principles discussed above contain some ways to implement them, some of the reviewed papers discuss overarching methods that can be used to implement responsibility as a whole to AI systems.

AI organisations looking to develop responsible AI systems should adopt a "learning governance model", where decision-making is based on reflection and overseen by stakeholders (Morley et al., 2021). They should distribute responsibility for their systems on the development teams working on them, and create processes to hold these teams responsible for unintended occurrences within their field of responsibility (Rakova et al., 2021). In addition, these organisations should consider creating multi-disciplinary advisory boards for AI ethics (Q. Lu et al., 2022; Y. Wang et al., 2020). Leaders of such organisations should work to show their commitment to responsible practices (Papagiannidis, Mikalef et al., 2022), while celebrating responsible successes (Q. Lu et al., 2022) and openly admitting failures (Rakova et al., 2021) along the way. Responsible practice should be included as part of the eval-

uation of employees, to increase the personal incentives for developing responsible systems (Rakova et al., 2021). Firms looking to automate tasks should "consider the welfare of their employees, and how their actions may violate the norms associated with various social contracts" (Wright & Schultz, 2018, p. 827).

Many papers highlight the need for stakeholder involvement in order to develop responsible systems. Buhmann and Fieseler (2021) call for using open-forum discussions, where information can be shared and all stakeholders, especially those at risk of negative consequences, can raise their concerns and present suggestions. Similarly, Dignum (2017, 2019) highlights the need for education of stakeholders, if they are to fully participate in discussions. Clarke (2019) and Havrda and Rakova (2020) argue that analysis of a wide range of stakeholders should form a basis for risk assessments of the system, with the intention of discovering risks that are important to mitigate. Similarly, Dignum (2021) argue that stakeholders should be involved in decisions when models "use human data, affect human beings or have other morally significant impacts" (p. 4). Gianni et al. (2022) calls for including the public when defining "AI's role for our future societies" (p. 14). Finally, Lukkien et al. (2021) argue that minorities and people from different cultural background should be included in not only AI development, but also AI research.

Q. Lu et al. (2022, p. 104) call for including responsible AI principles in agile software development methods, such as sprint planning, testing and code reviews (Dybå & Dingsøyr, 2008). While this requires some adaption, primarily in the form of implementing testing methods and metrics for measuring principle implementation, doing so ensures frequent testing, evaluation and monitoring of the responsibility of an AI system, thus increasing the probability of a successful adoption of responsibility (Jobin et al., 2019; Nauck, 2019; Rakova et al., 2021).

Non-functional requirements (NFR), such as usability, performance and security, are considered a critical component of many existing software development projects (see e.g., Chung and Do Prado Leite, 2009; Glinz, 2007; Khatter and Kalia, 2013). Q. Lu et al. (2022) argue that many responsible AI principles, such as Fairness, Non-maleficence and Justice, can be implemented in the form of NFRs. This allows developers to implement responsible AI principles in existing systems with minimal adoption, thus lowering the barrier for making AI systems responsible.

Some of the reviewed papers argue that national governments should do more to ensure responsibility of AI systems operating within their countries. Brand (2022) and Chen (2020) call for national AI registers, where organisations must register any AI-powered systems they develop. These systems should undergo impact as-

sessments, and the level of impact it may have on humans decides the requirements of the system (Brand, 2022). For example, a system with significant impact may be required to implement "safeguards", systems that report on any errors or imbalances in the AI system (Chen, 2020). Reports are then sent regularly to the government, to facilitate auditing of the systems (Brand, 2022; Chen, 2020). Floridi et al. (2018) argue that the IT infrastructure of the justice system should be improved, in order to enable auditing of AI systems in court.

4.4 RQ2 – Antecedents for responsible AI

The antecedents for responsible AI presented by the reviewed papers were clustered using an open coding-approach. This is described by Corbin and Strauss (1990) as a method suitable for creating new insights, by clustering qualitative data without pre-defined categories. This method led to three clusters of antecedents: Laws and regulations, Stakeholder demands and loss of reputation, and Internal motivation.

Laws and regulations Laws and regulations appear to be the most common antecedent mentioned among the reviewed papers. The European Union's General Data Protection Regulation (GDPR) is a European law enforcing privacy rights (Chen, 2020; Nevanperä et al., 2021; Werder et al., 2022), giving European stakeholders the right to control which personal data can be used for decisions in AI systems (Hacker & Passoth, 2022), and a right to have personal data fully erased, if they so want (Fjeld et al., 2020). GDPR Article 13 further provides European citizens a right to "meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" (European Commission, 2016, p. 41), if they are the subject of automatic processing. While it does not involve the core principles directly, GDPR operates as an antecedent for the principles of Privacy and Transparency, requiring them for any AI system operating within the European Union. In addition, both Product Liability Law and Contract and Tort Law requires developers targeting the European Union to provide Transparency if doing so reduces risk more than it reduces the utility of the system, and Banking Law requires AI systems used by banks to be Transparent (Hacker & Passoth, 2022).

In addition to the European-based GDPR, other laws may require similar principles. In the US, the Health Insurance Portability and Accountability Act (HIPAA) protects personal information in a healthcare context, thus working as an antecedent for Non-maleficence for AI systems operating in US healthcare (Werder et al., 2022).

Similar protections are required under the Act on the Protection of Personal Information (APPI) in Japan, thus increasing Non-maleficence for AI systems on the island (Y. Wang et al., 2020). In Norway, the protections of GDPR are provided by Personopplysningsloven (2018), providing the same benefits to responsibility. Other places may have local regulations that include "principles of standards and ethical considerations, such as auditing processes and algorithmic impact assessments" (Papagiannidis, Mikalef et al., 2022, p. 60-61), thus enabling several of the principles and implementation methods described in Section 4.3.

None of the laws mentioned above directly affect AI systems, instead placing limitations on data usage and Transparency. The European Union has, however, recently proposed an Artificial Intelligence Act (AIA), that specifically targets AI systems, banning systems with unacceptable risk and placing strict requirements on high-risk systems (European Commission, 2021). If AIA is implemented, stricter regulations are likely to be placed on AI systems operating in the European Union, thus increasing the requirements of responsibility affecting these systems (Hacker & Passoth, 2022).

There are multiple potential reasons why Laws and regulations are the most common antecedent among the reviewed papers. Rakova et al. (2021) describe organisations as reactive, thus only acting when they are forced to do so by external factors, such as binding laws. At the same time, breaking AI regulation comes with potentially large economic costs, with severe infringements potentially resulting in a fine of 4% of the organisation's worldwide annual revenue (Wolford, n.d.). This provides a quantitative measure of responsibility of AI systems, lowering the economic benefit of developing irresponsible models and justifying the additional resources required for implementing principles (Morley et al., 2021).

Although laws and regulations help increase the responsibility of AI systems, there are some who criticise the current state of AI regulations. Chen (2020) states that GDPR only protects personal data, and thus does not protect users when no personal data is used. Developers interviewed by Henriksen et al. (2021) are worried that principles for responsible AI may delay new regulations that could provide clearer and stricter regulations on AI systems. Similar worries are also noted by Hagendorff (2020). Finally, Floridi et al. (2018, p. 694) argue that "compliance with the law is merely necessary [...], but significantly insufficient," and that AI organisations therefore must do more than simply comply with existing regulations.

Stakeholder demands and loss of reputation Buhmann and Fieseler (2021) propose that stakeholders are essential for AI systems to operate, and that organ-

isations that struggle to explain their systems when key stakeholders demand such information, risk their reputation. Similar concerns of reputation are echoed by Eitel-Porter (2021) and Havrda and Rakova (2020), showing the need for Transparency in AI systems. This need is further strengthened by the hidden inner workings typical for AI systems scaring stakeholders from engaging with the technology (Merhi, 2022). Adding on this, Clarke (2019, p. 1) argue that "it is in the interest of all organisations to avoid their stakeholders suffering harm," and that such suffering can be only be minimized by making sure AI systems follow the principle of Non-maleficence. As reputation losses bring with them the potential of significant economic losses, this can, like with regulation, be quantified and used to justify the cost of adhering to responsible principles (Hagendorff, 2020).

When providing AI systems to professional organisations, such organisations may have their own codes of ethics, which requires responsibility from their suppliers (Anagnostou et al., 2022). In such cases, having an irresponsible AI system may lead to loss of sales. As markets are "visible, dynamic and viable," such requirements make responsibility in AI systems a "must-have requirement for success" (Gupta et al., 2021, p. 2). Similar expectations may, over time, become apparent also with non-company stakeholders and customers (Y. Wang et al., 2020).

Although several papers argue why stakeholder demands work as antecedents for responsible AI, Morley et al. (2021, p. 8) observe that while laws and regulations actually lead to changed design practices, "other benefits, such as those related to public reputation and consumer loyalty, seem to motivate public declarations of compliance with principles, but do not yet provide sufficient motivation for altering design behaviours or practice." Rakova et al. (2021) found that organisations are reactive, and thus are only driven to act when the risk of catastrophic media attention is too high.

Internal motivation Kumar et al. (2021) state that internal motivation, together with regulations, are the primary antecedents for responsible AI development. Henriksen et al. (2021) found that the AI developers they interviewed were motivated by standards and certifications, rather than principles. In a similar fashion, Rakova et al. (2021) found that most current work on responsible AI in organisations was driven by individual workers and their voluntarily sacrificed resources. As human nature is inherently to be selfless and do good (Ward, 2012), it is natural to assume that most AI workers want to do good for humanity and create responsible AI systems.

4.4.1 Barriers preventing responsible AI

To understand factors that prevent AI from becoming responsible, the same approach was conducted on barriers presented in the reviewed papers. Once again, three clusters were discovered – Abstract principles, Opaque systems and Toothless guidelines.

Abstract principles One core barrier presented by the reviewed papers is the abstractness of responsible AI principles, and lack of methods for implementing the details in practice (see e.g. Barredo Arrieta et al., 2020; Havrda and Rakova, 2020; Jakesch et al., 2022; Lukkien et al., 2021). Gianni et al. (2022) argue that there are "difficulties in translating principles and operational normative tools, such as transparency or fairness, into concrete measures" (p. 9), although, according to Miller and Coldicutt (2019; as cited by Morley et al., 2020), 78% of tech workers want practical tools for evaluate ethical impacts of their products. Morley et al. (2021) find that AI practitioners need assistance that "go beyond general guidelines for professional and ethical practice", and instead is "practically applicable to real-world ethical decisions" (p. 2). Multiple papers (e.g., Barredo Arrieta et al., 2020; Chen, 2020; Rakova et al., 2021) note that there is a gap between what is being researched in academia and what is actually implemented in organisations.

Opaque systems Another barrier comes as a result of AI technology often being proprietary, limiting the information organisations are willing to share about their system (Buhmann & Fieseler, 2021; Hacker & Passoth, 2022). Although calls have been made for sharing the data and source code used to develop systems, doing so may not lead to public understanding of how the system works, due to the opacity of prevalent AI methods (Buhmann & Fieseler, 2021; Siala & Wang, 2022). At the same time, increasing transparency may increase the risk of hacking, thus limiting the Non-maleficence of the system (Cheng et al., 2021).

Toothless guidelines One issue that ethical guidelines are considered toothless (Henriksen et al., 2021), and there is a lack of ways to easily enforce them (Hagendorff, 2020; Liu et al., 2021). As long as laws do not regulate AI systems in detail, this leaves stakeholders with few options for remedy against irresponsible AI systems (Henriksen et al., 2021). McNamara et al. (2018) (cited in Nevanperä et al., 2021) found almost no effect from ethical guidelines on the work of software developers. Similar results are found by Hagendorff (2020). This barrier is further strengthened

by companies who believe that following laws and regulations are sufficient for having ethical systems (Vakkuri et al., 2022).

4.5 RQ3 – Business advantages of responsible AI

To understand the advantages businesses may gain from making their AI systems responsible – and thus answer RQ3 – open coding was once again used on the outcomes and business advantages presented in the reviewed papers. This led to three clusters of business advantages – Increased reputation and stakeholder trust, Long-term profitability, and Internal improvements.

Increased reputation and stakeholder trust Most organisations looking to make their AI systems responsible primarily do so to minimize risks (Cheng et al., 2021; Eitel-Porter, 2021), as irresponsible use of AI can lead to severe harm to an organisation's reputation (Hagendorff, 2020; Siala & Wang, 2022). Implementing responsibility in AI systems means – by definition – that irresponsible use of AI is reduced, and can thus both protect responsible AI organisations from reputational harms (Buhmann & Fieseler, 2021), and increase their reputation among stakeholders (Werder et al., 2022).

Anagnostou et al. (2022), Gupta et al. (2021) and Merhi (2022) all discuss that AI systems fulfilling the principles of Transparency, where decisions can be explained and discussed, and Justice, where systems avoid bias, are likely to lead to increased trust from customers and stakeholders, as decisions are fair and understandable for those affected by them. Similarly, ensuring AI systems follow the principle of Transparency enables external reviews and audits (Hacker & Passoth, 2022; Rizinski et al., 2022), which again increases stakeholder trust in the system (Reinoso, 2021). This trust can be further strengthened by being able to test and directly compare AI systems against traditional solutions, as suggested via the principle of Quality benchmarking proposed by Hacker and Passoth (2022). Ensuring that AI systems are beneficent, and clearly showing the beneficence of the system, is likely to lead to improved acceptance, and thus use, of an AI system (Liu et al., 2021; van Bruxvoort & van Keulen, 2021). In a similar fashion, W. Wang et al. (2021) found that following the principles of Autonomy, Transparency, Justice, Beneficence and Non-maleficence increased use of, and satisfaction with, AI systems in a healthcare context.

Long-term profitability Buhmann and Fieseler (2021) argue that by engaging directly with a wide range of stakeholders, organisations can profit by quickly gath-

ering new information, such as changing interests or reactions of stakeholders to new systems, that can give them an advantage compared to organisations that lack direct access to their stakeholders. Similarly, Zaheer and Trkman (2017) (referenced in Liu et al., 2021) found that creating trust among stakeholders will increase their willingness to share information. Liu et al. (2021) and Morley et al. (2021) found that responsible AI methods, through creating trust, sets up long-term relationships with customers, thus increasing customer loyalty. As trust has an economic value (Minkkinen et al., 2021), increasing trust indirectly increases the profitability of a firm. Profitability can be further strengthened by responsible AI methods helping organisations avoid or minimize costly mistakes (Floridi et al., 2018; Rakova et al., 2021).

Interviews conducted by Kumar et al. (2021) found that responsible methods increased users' perceived value of AI systems, which again increases the value of the system, suggesting that responsible AI systems indirectly leads to increased market share for AI organisations. Other papers repeat the connections between responsible AI and market share (Cheng et al., 2021) or organisational performance (Werder et al., 2022). Implementing responsibility in AI systems may also increase the brand value among customers looking for ethical solutions to their problems (Minkkinen et al., 2021; Y. Wang et al., 2020).

Internal improvements Adopting responsible AI practices may lead to improvements within the organisations, that can create long-lasting business advantages. Widén-Wulff and Ginman (2004) (referenced in Liu et al., 2021) argue that trust, and other similar values, can lead to corporate well-being and innovativeness. A shortage of qualified developers may give organisations committed to responsible AI an edge when it comes to recruiting new talent (Papagiannidis, Mikalef et al., 2022). By interviewing Nordic companies working with AI, Papagiannidis, Mikalef et al. (2022, p. 65) found that adopting responsible AI practices positively influenced the knowledge management capabilities of an organisation, which again "positively influenced competitive performance."

5 Discussion

The aim of this paper was to create an understanding of how AI systems can be designed to be responsible, through a systematic literature review. The results of the review, as well as answers to RQ1, RQ2 and RQ3 are presented in Section 4. This

section will attempt to connect these results and answers to the overarching aim of the paper, and use them to create a framework for responsible AI development. In addition, some observations are discussed, alongside limitations of this study – and potential for future research.

5.1 Developing responsible AI systems

Most of the reviewed papers agree that AI systems can be made responsible by following a set of ethical principles. Following the categorization given by Ryan and Stahl (2021), this paper found 11 groups of principles for responsible AI. By adapting the notion of core principles used by Canca (2020) and Floridi et al. (2018), these were then distributed in two categories, resulting in four core and seven instrumental principles.

For an AI system to be responsible, it should strive to maximize the four core principles. First, it should be designed to provide Beneficence, by working towards doing good for humanity. It should ensure users and stakeholders maintain their Autonomy as far as possible, and that the decisions made represent Justice, in that they are not biased or discriminating. Finally, the AI systems should be Non-maleficent, in that it should do as much as possible to prevent harm, whether that is towards humans, other living things, the environment or the planet as a whole.

To achieve this, AI systems should be designed based on a set of instrumental principles. These provide responsibility through supporting and facilitating for the core principles, and should be adapted to the organisation developing and using the system, the context it is used in, and the goals and implementations of the system itself. An AI system could be Transparent, in that decisions can be explained and understood, which typically leads to increased Justice and Autonomy. The system could be Accountable, in that responsibility for its decisions are distributed among organisations and individuals creating and using the system, which facilitate the Justice and Non-maleficence of the system. The same principles can be strengthened by ensuring the Privacy of the users and stakeholders impacted by the system. Finally, systems can be Sustainable, in that their environmental footprint is minimized, or Trustworthy, by ensuring they work optimally and as expected, both contributing to increasing the core principles.

As shown in Section 4.4.1, a key barrier preventing a wide adoption of responsible AI is the gap between abstract ethical principles and the concrete work being done by AI developers in the industry. To aid this, this paper developed a framework for

selecting principles relevant for a project, and converting these to concrete implementations that can be followed, tested and used during development of the system. This principle is presented in Section 5.2.

5.2 The EPNIS framework

The EPNIS-framework is a project-level framework created to help AI developers, designers and managers assess a proposed project, discover relevant ethical principles for the project and convert the abstract principles into concrete implementations that can be used during development. The framework is based on the Design for Values-approach described by Dignum (2019), but places it in a more concrete and development-focused framework. An overview of the EPNIS framework is given in Figure 3.

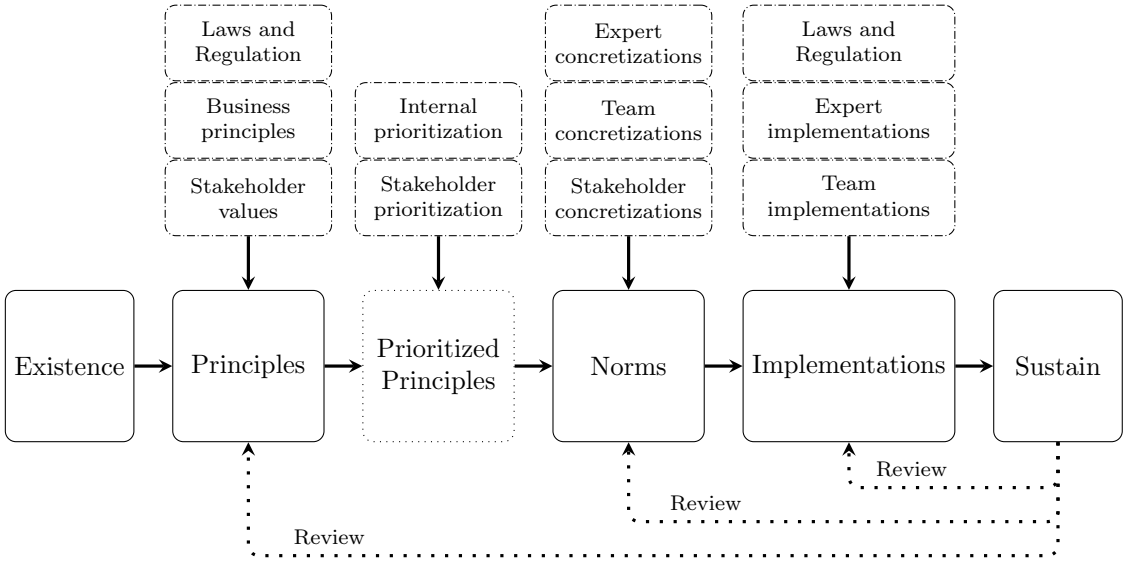

Figure 3: Overview of the EPNIS framework

As calls have been made among the reviewed papers to include stakeholders in the development of responsible AI, these are included in all relevant steps of the framework. While selection of stakeholders depend on the project, it is crucial that this include as wide range of people as possible, in order to ensure everyone affected by the AI system gets to share their opinion on how the project should be developed.

5.2.1 Steps of the EPNIS framework

The EPNIS-framework consists of four main steps – Existence, Principles, Norms, Implementations and Sustain (EPNIS), as well as one substep – Prioritized principles. These steps are explained in detail in the following pages.

Existence Every responsible AI project should start with assessing whether the system can be developed responsibly and should exist in the first place. This entails checking whether the goals of the system are beneficial and non-maleficent, and that these can be achieved without negatively impacting autonomy and justice in any significant way. If this is not possible, then the system cannot be developed in a responsible way, and should be abandoned.

Principles Once it is clear that the project can be developed responsibly, the process moves to the second step of the framework – Principles. Here, the development team should use relevant laws and regulations for their system and context, together with existing business principles for their organisation and stakeholder values, to select ethical principles that apply to the project. It is important to include a wide range of stakeholders in this step, to ensure that all possibly relevant principles are considered for the system.

As discussed in Section 4.3, core principles should be part of any responsible AI project, as they are the ones that directly contribute to the ethical value of a project. As such, the goal of this stage is to find the instrumental principles that contribute the most to the core principles of Autonomy, Beneficence, Non-maleficence and Justice.

Prioritized principles Once the project has a set of principles, the development team should combine internal prioritization of principles (from organisational strategies, organisation values, ethics panels and internal to the team) with stakeholders' prioritization, to create a list of prioritized principles. While ethical dilemmas may still appear, driving ethical discussion with no right or wrong answer, this gives the development team a starting point for the next steps, while also operating as a prioritization of limited resources.

Norms After prioritizing principles, the development team should utilize expert panels, stakeholders and the team itself to concretize the principles into norms, in the same way it should be done in the Design for Values-approach used by Dignum (2019). These norms are normative interpretations of the principles, and primarily state how the principles should be interpreted.

An example of converting a principle to a norm is shown by Benjamins et al. (2019), who describes three different interpretations of the principle of fairness (included in the core principle of Justice): Independence, Separation and Sufficiency. To better

show the differences, each interpretation will be illustrated with an AI system meant to approve loans, where gender is a sensitive variable. Positive predictions means that the system believes the applicant will pay back the full loan, and a true positive means it is correct in its beliefs.

Independence means that the proportion of positive predictions are the same for all sensitive groups (Benjamins et al., 2019). In the example, this means that if half of all men are approved for loans, half of all women should also be approved for loans.

Separation means that the true positive and false positive rate is the same in all sensitive groups (Benjamins et al., 2019). In the example, this means that if 80% of all men that end up paying their loans are approved by the system, 80% of all women that end up paying their loans should also be approved. At the same time, if 90% of men who fail to pay their loans are not approved, then 90% of women in the same situation should also not be approved.

Sufficiency means that the Positive Predictive Value (i.e, the proportion of true positive predictions divided by all positive predictions) is the same for all sensitive groups (Benjamins et al., 2019). In the example, this means that if 95% of men who are approved for loans end up paying them back, then 95% of women who are approved for loans should end up paying them back.

As it is impossible to achieve all three, the development team for this hypothetical scenario would have to decide, using all available resources, which interpretation suits their context, organisation, system and personal beliefs the best.

Implementations When the norms of the project are in place, the development team should use relevant laws and regulations, implementations used in academia and by experts, as well as their own ideas to create concrete implementations of the selected norms. These implementations should ideally be testable and measurable, and should, following Design for Values (Dignum, 2019), be explicitly connected to a specific part of the system.

By creating concrete and testable implementations, these can be included in common software development practices, such as sprint plannings and code reviews (Dybå & Dingsøyr, 2008). Each implementation is connected to a certain part of a system, making auditing of the implementation easy, as only specific parts of the system needs to be checked. As there is a direct line from principles, through norms

to implementations, such audits also check that the norms, and by extension the principles, still apply to the system.

Sustain After all implementations have been designed, development of the AI system can begin. It is important that the developers keep the previously selected implementations in mind as they work on the system, to avoid accidentally disabling an implementation, thus disconnecting a norm from the system. This is made easier by connecting each implementation to specific parts of the system, as developers working on one part only need to remember the implementations connected to their area of work.

Modern software development is a flexible process (Dybå & Dingsøyr, 2008), and ethical values and bias may change over time (Nauck, 2019). This makes it important to review the previous steps of the framework, and managers of AI development projects should ensure regular reviews are conducted.

Implementations are most closely related to the actual development process, and will shift as the project is developed. These should therefore be reviewed most frequently, to ensure that they are still implemented and relevant as the AI system adapts throughout its development. For AI organisations utilizing agile development methods, these reviews could be conducted as part of code- or sprint reviews (Dybå & Dingsøyr, 2008).

Norms only shift when the interpretation of principles shift. As such shifts may occur due to changes in the context or organisation, changes in the goals or design of the system, or through larger societal changes, these should still be reviewed regularly, but less frequently than the implementations. While the actual frequency depends on the specifics of the context, organisation and system, an example is to review these on a biannual or quarterly basis. If any norms are changed or adjusted during these reviews, implementations should be reviewed again, to ensure that the connection between principles, norms and implementations still hold.

Principles are the least likely to shift, as this only occurs if laws and regulations, organisational values or societal values change. As such, these reviews may be conducted on a reactive basis, such as when relevant laws are changed, or on an annual, semi-annual or less frequent basis. If any principles, or the prioritization of these, are changed during the review, norms and implementations should also be reviewed again, to ensure that the connection between principles, norms and implementations still hold.

During the process of implementing and following the EPNIS framework, any de-

cisions taken along the way and the results of each step, as well as the reviews, should be clearly documented and stored. To provide transparency and auditability to the system, these documents should be made available for audits when needed, and should thus be understandable by both internal and external auditing teams.

5.2.2 The EPNIS framework compared to existing frameworks

Some of the reviewed papers introduce frameworks for developing responsible AI. While the EPNIS framework has been designed to solve a previously unsolved gap between principles and practice, it overlaps with these existing frameworks on some accounts, and a comparison is therefore in order.

Chen (2020) introduce a framework for AI governance, built around a national registry of AI systems. Their framework is focused on regulating AI systems, and therefore applies to AI organisations and government systems, rather than project-level implementations of principles. The EPNIS framework can be used as a safeguard for high-impact systems, can provide documentation that can be used for creating annual reports, and can be used as a step towards increased auditability, all things that are included as part of the framework introduced by Chen (2020).

Dignum (2019) provides the Design for Values framework. As this is used as foundation for the EPNIS framework, the two align with each other, most notably through the connection between principles, norms and implementations. Although the frameworks are somewhat alike, the EPNIS framework focuses on the implementation details, including relevant inputs that should be used for each step, whereas Design for Values primarily attempts to connect existing software development methodologies to responsible AI. Instead of directly comparing the two frameworks, then, they should be used alongside each other, using Design for Values to find principles, norms and implementations, and using EPNIS to systematize the process, find the right inputs for each step and ensure steps are reviewed when needed.

AI4People, the framework introduced by Floridi et al. (2018), is mostly focused on how policy makers may facilitate for responsible AI. The two frameworks therefore do not compete with each other. Instead, EPNIS can be used to aid some of the action points proposed in AI4People, such as by facilitating for auditing through documentation of each step.

Peters et al. (2020) present two frameworks for responsible AI – the Responsible Design Process (RDP), and the Spheres of Technology. The Responsible Design Process consists of five steps showcasing different development stages, with Re-

search, Insights, Ideation, Prototypes and Evaluation. Although the framework was originally developed to responsibly design new technology, it has great potential for working alongside the EPNIS framework. The first two stages, Research and Insight, are used to discover needs and preferences from stakeholders. These can be used to either support, or fully conduct, step two and three of EPNIS – Principles and Norms – where the goal is to combine different inputs, including from stakeholders, to create principles and norms for the system. The next two steps of the Responsible Design Process, Ideation and Prototypes, take previously generated insights and create solutions based on them, and can therefore be used to generate implementations from the previously discovered norms. Finally, the Responsible Design Process's Evaluation and EPNIS's Sustain have the same goal, continually reassessing the responsibility of a system. This shows that the two frameworks – the Responsible Design Process and EPNIS – have great potential for success when used together.

The second framework introduced by Peters et al. (2020), the Spheres of Technology, maps the impact of technology to six different spheres – Adoption, Interface, Task, Behavior, Life, and Society. This is differs from EPNIS in multiple ways, but can be used to ensure the selected principles, norms and implementations cover all possible impacts of the AI system.

Werder et al. (2022) introduce a framework for mitigating bias, thus supporting Justice, through data governance. This may be used as an implementation for responsible AI systems developed using the EPNIS framework, if it is relevant to the context, organisation and design of the system being developed.

The framework introduced by Wright and Schultz (2018) is designed to gather expectations and inputs from stakeholders, and evaluating how well these expectations are met. This could integrate with the EPNIS framework in two ways. First, the framework from Wright and Schultz (2018) could be used during the Existence-step of the EPNIS framework, to evaluate whether the stakeholders' expectations can realistically be met in a responsible way, and thus whether the system should exist in the first place. Secondly, the expectations can be used to create principles and norms when following the EPNIS framework, as they give a good indication of stakeholder values, prioritizations and concretizations. The two frameworks, then, should be used alongside each other to develop responsible AI systems.

Some additional papers also introduce what they refer to as frameworks for responsible AI. These are, however, targeting laws and regulations (Brand, 2022) or the finance industry (Rizinski et al., 2022), or focused on defining responsible AI's place

in society (Buhmann & Fieseler, 2021), categorizing algorithms (Cheng et al., 2021), assessing the impact of AI (Havrda & Rakova, 2020), mapping expectations of socio-technical systems (Minkkinen et al., 2021) or evaluating how organisational culture impacts responsible AI initiatives (Rakova et al., 2021), rather than finding ways to implement responsible AI. As such, comparing these to the EPNIS framework is not relevant for this paper.

5.3 Observations

As shown in Table 5, only five of the reviewed papers (9%) conducted quantitative surveys, and only four (7%) conducted qualitative surveys. Of these, only five papers included surveys conducted on stakeholders not working directly on developing AI (i.e., representative samples of the public, healthcare professionals, etc.). This shows that most of the reviewed papers look inward, at previous research, existing guidelines, experts, and professionals developing AI systems. While basing work on existing research is important (Tranfield et al., 2003), this prevents new inputs from outsiders, and could potentially lead to interesting insights from stakeholders being missed. Similar exclusion of stakeholders has also been found in commercial guidelines for responsible AI, where only 38% of the papers reviewed by Bélisle-Pipon et al. (2022) engaged stakeholders – and only 9% involved ordinary citizens. A worst-case scenario is that stakeholders are left out of the process of defining what constitutes a responsible and ethical AI system. As the goal of responsible AI is to develop AI systems that operate for the good of humanity, this could ultimately lead to a collapse of the whole research field.

The geographical distribution, as shown in Table 6, show a bias towards Western research, and thus the potential for a bias in the reviewed principles. This geographical bias is reflected in the distribution of guideline documents that are available. AlgorithmWatch (n.d.) contains a set of 167 guidelines for developing responsible AI. While this is not necessarily an exhaustive list, the database has been in use since 2019 and is open for submissions (AlgorithmWatch, 2020), making it a representative source. Of the 167 guidelines in the database, only 14 (8%) are from Asian countries, 1 (0.6%) from Southern Africa and no guidelines come from South America or Northern Africa. As mentioned in Section 4.3.2, ethical values may change depending on the geographical background of relevant stakeholders. As such, this bias may limit the generalizability of the current state of responsible AI, especially for AI systems in Asian, South American or African contexts.

Both responsible AI and RRI (as discussed in Section 2) have the same goal –

developing responsible solutions to societal problems. The responsible AI principles discussed in this paper are quite different from the conceptual dimensions used by RRI, as the AI principles primarily focus on the outcomes of the system, while RRI primarily focus on the innovation process. Still, the methodology them employ are still the same – both strategies involve stakeholders and base their development process on a set of principles, with the goal of achieving responsibility. These same observations have also been noted among the reviewed papers. Hacker and Passoth (2022, p. 346) mention that responsible AI is "similar, but still quite disconnected" from RRI. Similarly, Dignum (2019) argue that RRI can be applied to responsible AI development. Based on this, AI developers and managers should adopt methods and tools from RRI, with the goal of achieving responsibility in both the resulting system, as well as the research and innovation processes leading up to the finished system.

In a time when countries are accused of using sports for sportswashing (Fruh et al., 2022; Wearing, 2022), organisations are accused of using fake eco-friendliness for greenwashing (Gibbens, 2022), and even the US military is accused of using Pride for pinkwashing (Kane, 2022), some (e.g. Gianni et al., 2022; Havrda and Rakova, 2020; Morley et al., 2021) have raised concerns that responsible AI can be used for ethics washing, i.e., that organisations may publicly adopt ethical principles to show that their AI systems are responsible and ethical, without changing the underlying, unethical practice (Morley et al., 2021). These organisations would then claim benefits of responsible AI, such as increased stakeholder trust and long-term profitability, without actually having responsible systems. To avoid this, organisations should set long-term responsible goals, and publicly show their processes towards responsibility, in addition to the overarching principles they have adopted (Rakova et al., 2021). By being transparent about how responsible AI is achieved, not just the abstract principles constituting the goal, organisations can ensure their stakeholders that the AI systems they interact with are actually safe, thus sparing them from such accusations.

An interesting observation is that trust is mentioned as one of the core advantages of developing responsible AI, yet it is one of the least popular principles among the reviewed papers. Although pinpointing the exact reason for this is hard, a potential reason is that trust requires two parties, a trustor, i.e., the party that trusts, and a trustee, i.e., the party that is trusted (Huang & Fox, 2006). In comparison, the other principles, such as beneficence, only require one party, the party that creates benefits. This means that trust is not something that can be created internally in an organisation or AI system, but must be created in partnership with trustors. By using stakeholders as trustors, it is clear that trust is something that must be

created through having trustworthy systems, and should therefore be considered as the output of responsible AI, rather than the input.

5.4 Limitations

In the words of Morley et al. (2020, p. 2160), "all research projects have their limitations and this one is no exception." Although this paper attempts to create an understanding of how AI systems can be responsible, it does so by solely looking at academic papers. As shown by AlgorithmWatch (n.d.), much work has been done on developing guidelines for responsible AI in both governments and industries around the world. Although some of the reviewed papers have reviewed these (e.g., Bélisle-Pipon et al., 2022; Fjeld et al., 2020; Jobin et al., 2019; Ryan and Stahl, 2021) excluding non-academic guidelines from the review brings with it the potential for missing several important observations, principles and frameworks that are not included in the reviewed papers.

Although the methodology of systematic literature reviews is designed to minimize bias, Tranfield et al. (2003) highlights the eligibility assessment of papers to be a potential source for bias, as subjective opinion is needed to evaluate whether a given document should be included in the review. This is especially true for this paper, as all the research has been conducted by one person. This leaves limited possibility to discuss any uncertainties, although my supervisor has available for some cases. Although this creates a risk that the selection of papers – and thus the included principles – may be biased, the fact that the found principles are typically used in multiple reviewed papers indicates that the results of this paper are likely to be relatively generalizable.

AI ethics is a wide field. A key limitation of this research is that it solely focused on one type of ethical AI – responsible AI. Several of the reviewed papers mention more of these types, and Table 9 shows some of these, together with the number of articles using the term on Scopus. Interestingly, several of the included types have more results than responsible AI. This shows that while including multiple types would increase the scope of the paper, doing so would also provide new perspectives on how to develop AI responsibly, with the potential of uncovering new principles, norms and implementations, as well as frameworks and tools to convert these to practice.

Table 9: Search results for types of ethical AI. Searches were conducted on December 14, 2022, using the search string `"synonym AI" OR "synonym Artificial Intelligence"`, for example, `"Responsible AI" OR "Responsible Artificial Intelligence"`. Searches looked at article titles, abstracts and keywords.

Synonym	**Scopus results**
Explainable AI	4260
AI ethics	620
Trustworthy AI	356
Ethical AI	267
Human-centred AI	245
Responsible AI	237
AI for social good	74
Sustainable AI	48
Fair AI	42

5.5 Future research

Ethical AI included several, somewhat overlapping terms, as shown in Table 9. To unify this wide range of terms, future research should consider creating a complete taxonomy for the field of responsible AI. As taxonomies "help humans classify objects according to similarities and differences" (Kundisch et al., 2022) and formally create a unifying definition of terms (Uschold & Gruninger, 1996), this could create a set of agreed-upon terms, with clearly defined separations, that would allow research to move in the same direction. Although this has not yet been done for the field of AI ethics, significant efforts have gone into creating taxonomies of explainable AI (e.g., Barredo Arrieta et al., 2020; Bellucci et al., 2021; Brennen, 2020; Clinciu and Hastie, 2019; Graziani et al., 2022), and these could be used as a starting point for mapping the whole field of AI ethics.

Section 4.2.1 looks at different definitions of artificial intelligence used throughout the reviewed papers. While a deep-dive into definitions and terminology is outside of the scope of this paper, there appears to be a lack of consensus regarding how to define the concept of artificial intelligence, which should be considered the backbone of responsible AI. As such, AI researchers should work together with researchers within the field of philosophy and linguistics to create a unifying, agreed-upon definition of artificial intelligence. This, much like the above-mentioned taxonomy, would align future research along agreed-upon lines, thus ensuring work is moving the field in the same direction, while limiting disagreements and confusion.

As shown in Figure 2, the most used principles among the reviewed papers are Transparency and Justice. Hagendorff (2020, p. 103) claim that this is because

these principles reflect "the 'male way' of thinking about ethical problems", in that they are focused on, and solvable by, technical tools and methods. To increase the adoption of the other principles discussed in Section 4.3, especially the remaining core principles of Autonomy, Beneficence and Non-maleficence, future research should work on developing tools and methods for facilitating, measuring and evaluating these. Two benefits would arise from being able to measure and evaluate the remaining principles. First, this would give AI developers and managers a better view of the responsibility of their systems, and would provide them with tools to ensure a targeted level of responsibility is achieved. Secondly, this would give lawmakers and governments an avenue for creating new laws that can ensure the responsibility of AI systems. As laws and regulations are major antecedents for responsible AI development, this would likely lead to an increase in the responsibility of both existing and new AI systems.

While the goal of this paper is to understand how to develop responsible AI systems, little attention has been paid to AI governance, or organisational governance as a whole. While the introduced framework is designed to push AI development in a responsible direction, this only makes up a small part of a much wider field of AI governance (Papagiannidis, Enholm et al., 2022). Gianni et al. (2022) discuss some methods for AI governance, but note that existing methods are limited. Future research should look at ways to properly govern AI systems in order to ensure they are responsible, and should consider how the EPNIS frameworks fit with existing, and future, methods for doing so.

Although the EPNIS framework is proposed to bridge the gap between abstract principles and actual development, it remains to be tested in real scenarios. Work should therefore be done to empirically evaluate the proposed framework, with the goal of both testing that it helps bridge the gap, as well as to discover ways the framework should be expanded to better fulfill this goal. This work can be done by conducting case studies, thus actually develop AI systems using the framework; by qualitative studies, such as surveying AI developers and -professionals to evaluate whether the framework could be applicable, useful and relevant for their work; or by quantitative studies, such as surveying AI organisations to estimate the probability of the framework being able to help bridge the gap for a given organisation.

Although several studies call for the inclusion of stakeholders in responsible AI development, few of the reviewed papers survey stakeholders. Future research should therefore be conducted with a focus on stakeholders of AI systems. This would could take several forms. Some work should attempt to map stakeholders for a given AI system, thus making it easier for developers to discover relevant stakeholders for

their work. These kind of stakeholder maps have been created in several other domains, such as fossil fuel (Yudha et al., 2018), digital health (An et al., 2022) and nanotechnology (Hansen, 2010), so this work only needs to apply existing methods to the field of responsible AI. Other work could aim directly at the general public, to learn what principles and norms are important to laypersons. This work should first employ qualitative methods to get an overview of all potentially relevant principles and norms, before using quantitative methods to gain an understanding of which principles and norms are important to the general public. Although Jakesch et al. (2022) has already conducted one such survey, their work is based solely in the US, so more work is needed to be able to generalize the results.

The EPNIS framework argues that organisations looking to develop responsible AI systems should find relevant principles, norms and implementations for their context, organisation and system to be developed. While the framework mentions inputs that should be considered, the actual process of finding these principles, norms and implementations is not described in the framework, nor in the rest of the paper. Although a selection of principles are described in Section 4.3, highlighting that every AI system should focus on core principles, no methodology is provided for evaluating potential instrumental principles for a given system. As the selection process will vary depending on the context of a system, future research should apply the framework to different contexts and systems, with the goal of finding similarities in both processes and results that can be generalized. These findings should be used to both expand the current knowledge on how to assess principles for responsible AI, as well as to extend the EPNIS framework with new knowledge and tools.

As there have been complaints regarding the barrier between academic research and responsible AI work in the industry, future research should adopt an implementation-based view, and prioritize ways to concretize and implement their findings in real-life scenarios.

6 Conclusion

This paper looked at the current state of responsible AI research, by conducting a systematic literature review. The review resulted in a set of four core principles – Autonomy, Beneficence, Non-maleficence and Justice – that create responsible value by themselves, and seven instrumental principles – Transparency, Accountability, Trust, Sustainability, Privacy and Others – that create responsible value by facilitating for and supporting the core values. Additionally, the reviewed literature

included three clusters of antecedents for responsible AI, as well as three clusters of business advantages that can be gained by adopting responsible AI practices.

These findings were used to create a project-level framework for responsible AI development, the EPNIS framework, designed to bridge the gap between abstract ethical principles and actual AI development. This framework is compared to existing frameworks for responsible AI, to show how it fits into the ecosystem of responsible AI development. Finally, the findings are used to point out future research that is needed within the field of responsible AI.

7 References

7.1 Literature review

Anagnostou, M., Karvounidou, O., Katritzidaki, C., Kechagia, C., Melidou, K., Mpeza, E., Konstantinidis, I., Kapantai, E., Berberidis, C., Magnisalis, I., & Peristeras, V. (2022). Characteristics and challenges in the industries towards responsible AI: A systematic literature review. *Ethics and Information Technology*, *24*(3), 37. https://doi.org/10.1007/s10676-022-09634-1

Balagué, C. (2021). The challenge of responsible AI. In M. Pagani & R. Champion (Eds.), *Artificial intelligence for sustainable value creation* (pp. 99–121). Edward Elgar Publishing Ltd. https://doi.org/10.4337/9781839104398

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bélisle-Pipon, J., Monteferrante, E., Roy, M., & Couture, V. (2022). Artificial intelligence ethics has a black box problem. *AI and Society*. https://doi.org/10.1007/s00146-021-01380-0

Benjamins, R., Barbado, A., & Sierra, D. (2019). Responsible AI by design in practice. https://doi.org/10.48550/ARXIV.1909.12838

Borda, A., Molnar, A., Neesham, C., & Kostkova, P. (2022). Ethical issues in AI-enabled disease surveillance: Perspectives from global health. *Applied Sciences (Switzerland)*, *12*(8). https://doi.org/10.3390/app12083890

Brand, D. (2022). Responsible artificial intelligence in government: Development of a legal framework for South Africa. *JeDEM - eJournal of eDemocracy and Open Government*, *14*(1), 130–150. https://doi.org/10.29379/jedem.v14i1.678

Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, *64*. https://doi.org/10.1016/j.techsoc.2020.101475

Canca, C. (2020). Operationalizing AI ethics principles. *Commun. ACM*, *63*(12), 18–21. https://doi.org/10.1145/3430368

Chen, L. (2020). A conceptual framework for AI system development and sustainable social equality. *2020 IEEE / ITU International Conference on Artificial Intelligence for Good, AI4G 2020*, 101–106. https://doi.org/10.1109/AI4G50087.2020.9310984

Cheng, L., Varshney, K., & Liu, H. (2021). Socially responsible AI algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, *71*, 1137–1181. https://doi.org/10.1613/JAIR.1.12814

Clarke, R. (2019). Principles and business processes for responsible AI. *Computer Law and Security Review*, *35*(4), 410–422. https://doi.org/10.1016/j.clsr.2019.04.007

Dignum, V. (2017). Responsible artificial intelligence: Designing AI for human values. *ITU Journal*. https://www.itu.int/en/journal/001/Pages/01.aspx

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way* (1st ed.). Springer Cham. https://doi.org/10.1007/978-3-030-30371-6

Dignum, V. (2021). The role and challenges of education for responsible AI. *London Review of Education*, *19*(1), 1–11. https://doi.org/10.14324/LRE.19.1.01

Doorn, N. (2021). Artificial intelligence in the water domain: Opportunities for responsible use. *Science of the Total Environment*, *755*. https://doi.org/10.1016/j.scitotenv.2020.142561

Eitel-Porter, R. (2021). Beyond the promise: Implementing ethical AI. *AI and Ethics*, *1*(1), 73–80. https://doi.org/10.1007/s43681-020-00011-6

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI* (tech. rep.). http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, *1*(1). https://hdsr.mitpress.mit.edu/pub/l0jsh9d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—an ethical framework for a Good AI Society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gianni, R., Lehtinen, S., & Nieminen, M. (2022). Governance of responsible AI: From ethical guidelines to cooperative policies. *Frontiers in Computer Science*, *4*. https://doi.org/10.3389/fcomp.2022.873437

Gupta, S., Kamboj, S., & Bag, S. (2021). Role of risks in the development of responsible artificial intelligence in the digital healthcare domain. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10174-0

Hacker, P., & Passoth, J. (2022). Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond (A. Holzinger, R. Goebel, R. Fong, T. Moon, K. Müller & W. Samek, Eds.). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13200 LNAI*, 343–373. https://doi.org/10.1007/978-3-031-04083-2_17

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Havrda, M., & Rakova, B. (2020). Enhanced well-being assessment as basis for the practical implementation of ethical and rights-based normative principles for AI. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2754–2761. https://doi.org/10.1109/SMC42975.2020.9283137

Henriksen, A., Enni, S., & Bechmann, A. (2021). Situated accountability: Ethical principles, certification standards, and explanation methods in applied AI. *AIES 2021 - Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 574–585. https://doi.org/10.1145/3461702.3462564

Jakesch, M., Buçinca, Z., Amershi, S., & Olteanu, A. (2022). How different groups prioritize ethical values for responsible AI. *ACM International Conference Proceeding Series*, 310–323. https://doi.org/10.1145/3531146.3533097

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kumar, P., Dwivedi, Y. K., & Anand, A. (2021). Responsible artificial intelligence (AI) for value formation and market performance in healthcare: The mediating role of patient's cognitive engagement. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10136-6

Liu, R., Gupta, S., & Patel, P. (2021). The application of the principles of responsible AI on social media marketing for digital health. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10191-z

Lu, Q., Zhu, L., Xu, X., Whittle, J., & Xing, Z. (2022). Towards a roadmap on software engineering for responsible AI. *Proceedings - 1st International Conference on AI Engineering - Software Engineering for AI, CAIN 2022*, 101–112. https://doi.org/10.1145/3522664.3528607

Lukkien, D., Nap, H., Buimer, H., Peine, A., Boon, W., Ket, J., Minkman, M., & Moors, E. (2021). Toward responsible artificial intelligence in long-term care: A scoping review on practical approaches. *The Gerontologist*. https://doi.org/10.1093/geront/gnab180

Merhi, M. (2022). An assessment of the barriers impacting responsible artificial intelligence. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-022-10276-3

Mikalef, P., Conboy, K., Lundström, J., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems*, *31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Minkkinen, M., Zimmer, M., & Mäntymäki, M. (2021). Towards ecosystems for responsible AI: Expectations on sociotechnical systems, agendas, and networks in EU documents (D. Dennehy, A. Griva, N. Pouloudi, Y. Dwivedi, I. Pappas & M. Mäntymäki, Eds.). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12896 LNCS*, 220–232. https://doi.org/10.1007/978-3-030-85447-8_20

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising AI ethics: Barriers, enablers and next steps. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01308-8

Nauck, D. (2019). Responsible AI. *Journal of the Institute of Telecommunications Professionals*, *13*, 14–19.

Nevanperä, M., Helin, J., & Rajamäki, J. (2021). Comparison of European Commission's ethical guidelines for AI to other organizational ethical guidelines. *Proceedings of the 3rd European Conference on the Impact of Artificial Intelligence and Robotics ECIAIR 2021*, 130–137.

Papagiannidis, E., Mikalef, P., Krogstie, J., & Conboy, K. (2022). From responsible AI governance to competitive performance: The mediating role of knowledge management capabilities (S. Papagiannidis, E. Alamanos, S. Gupta, Y. Dwivedi, Y. Dwivedi, M. Mäntymäki & I. Pappas, Eds.). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13454 LNCS*, 58–69. https://doi.org/10.1007/978-3-031-15342-6_5

Peters, D., Vold, K., Robinson, D., & Calvo, R. (2020). Responsible AI: Two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, *1*(1), 34–47. https://doi.org/10.1109/TTS.2020.2974991

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1). https://doi.org/10.1145/3449081

Rizinski, M., Peshov, H., Mishev, K., Chitkushev, L., Vodenska, I., & Trajanov, D. (2022). Ethically responsible machine learning in fintech. *IEEE Access*, *10*, 97531–97554. https://doi.org/10.1109/ACCESS.2022.3202889

Rothenberger, L., Fabian, B., & Arunov, E. (2019). Relevance of ethical guidelines for artificial intelligence: A survey and evaluation. https://www.researchgate.net/publication/332441075_Relevance_of_Ethical_Guidelines_for_Artificial_Intelligence_-_A_Survey_and_Evaluation

Ryan, M., & Stahl, B. (2021). Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, *19*(1), 61–86. https://doi.org/10.1108/JICES-12-2019-0138

Siala, H., & Wang, Y. (2022). SHIFTing artificial intelligence to be responsible in healthcare: A systematic review. *Social Science and Medicine*, *296*. https://doi.org/10.1016/j.socscimed.2022.114782

Thelisson, E. (2018). Towards a computational sustainability for AI/ML to foster responsibility. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 386–387. https://doi.org/10.1145/3278721.3278784

Vakkuri, V., Kemell, K., Tolvanen, J., Jantunen, M., Halme, E., & Abrahamsson, P. (2022). How do software companies deal with artificial intelligence ethics? A gap analysis. *ACM International Conference Proceeding Series*, 100–109. https://doi.org/10.1145/3530019.3530030

van Bruxvoort, X., & van Keulen, M. (2021). Framework for assessing ethical aspects of algorithms and their encompassing socio-technical system. *Applied Sciences (Switzerland)*, *11*(23). https://doi.org/10.3390/app112311187

Vetrò, A., Santangelo, A., Beretta, E., & De Martin, J. (2019). AI: From rational agents to socially responsible agents. *Digital Policy, Regulation and Governance*, *21*(3), 291–304. https://doi.org/10.1108/DPRG-08-2018-0049

Wang, W., Chen, L., Xiong, M., & Wang, Y. (2021). Accelerating AI adoption with responsible AI signals and employee engagement mechanisms in health care. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10154-4

Wang, Y., Xiong, M., & Olya, H. (2020). Toward an understanding of responsible artificial intelligence practices. In T. Bui (Ed.), *Proceedings of the annual Hawaii International Conference on System Sciences* (pp. 4962–4971). IEEE Computer Society.

Werder, K., Ramesh, B., & Zhang, R. (2022). Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems*, *13*(2). https://doi.org/10.1145/3503488

Wright, S., & Schultz, A. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons*, *61*(6), 823–832. https://doi.org/10.1016/j.bushor.2018.07.001

7.2 References not used in literature review

Abeygunawardena, P., Vyas, Y., Knill, P., Foy, T., Harrold, M., Steele, P., Tanner, T., Hirsch, D., Oosterman, M., Rooimans, J., Debois, M., Lamin, M., Liptow, M., Mausolf, E., Verheyen, R., Agrawala, S., Caspary, G., Paris, R., Kashyap, A., ... Sperling, F. (2002). Poverty and climate change: Reducing the vulnerability of the poor through adaptation. https://www.oecd.org/env/cc/2502872.pdf

Accenture. (n.d.). *Artificial intelligence services*. Retrieved November 29, 2022, from https://www.accenture.com/us-en/services/ai-artificial-intelligence-index

Agrawal, M. (2021). The possibilities of AI in 2030: Transformation across dimensions. *Forbes*. Retrieved December 12, 2022, from https://www.forbes.com/sites/forbesbusinesscouncil/2021/08/23/the-possibilities-of-ai-in-2030-transformation-across-dimensions/?sh=5c0cff26b67a

AI for Good. (n.d.). *Autonomous drones saving lives and powering disaster preparedness*. Retrieved December 12, 2022, from https://aiforgood.itu.int/autonomous-drones-saving-lives-and-powering-disaster-preparedness/

Airbus. (n.d.). *Artificial intelligence: Capitalising on the value of data*. Retrieved December 12, 2022, from https://www.airbus.com/en/innovation/industry-4-0/artificial-intelligence

AlgorithmWatch. (n.d.). AI Ethics Guidelines Global Inventory. Retrieved December 8, 2022, from https://inventory.algorithmwatch.org/

AlgorithmWatch. (2020). About - AI Ethics Guidelines Global Inventory. Retrieved December 8, 2022, from https://inventory.algorithmwatch.org/about

An, Q., Kelley, M., & Yen, P. (2022). Stakeholder mapping on the development of digital health interventions for self-management among patients with chronic obstructive pulmonary disease in China (P. Otero, P. Scott, S. Martin & E.

Huesing, Eds.). *Studies in Health Technology and Informatics*, *290*, 1106–1107. https://doi.org/10.3233/SHTI220290

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. Retrieved December 12, 2022, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Autonomy. (n.d.). Merriam-Webster. Retrieved December 10, 2022, from https://www.merriam-webster.com/dictionary/autonomy

Azaria, A., Ekblaw, A., Vieira, T., & Lippman, A. (2016). MedRec: Using blockchain for medical data access and permission management. In M. Younas & I. Awan (Eds.), *Proceedings - 2016 2nd international conference on open and big data, obd 2016* (pp. 25–30). Institute of Electrical; Electronics Engineers Inc. https://doi.org/10.1109/OBD.2016.11

Barr, A. (2015). Google mistakenly tags black people as 'gorillas,' showing limits of algorithms. *The Wall Street Journal*. Retrieved December 12, 2022, from https://www.wsj.com/articles/BL-DGB-42522

Basile, A., Yahi, A., & Tatonetti, N. (2019). Artificial intelligence for drug toxicity and safety. *Trends in Pharmacological Sciences*, *40*(9), 624–635. https://doi.org/10.1016/j.tips.2019.07.005

Beauchamp, T., & Childress, J. (2001). *Principles of biomedical ethics*. Oxford University Press.

Bellucci, M., Delestre, N., Malandain, N., & Zanni-Merk, C. (2021). Towards a terminology for a fully contextualized XAI. In J. Watrobski, W. Salabun, C. Toro, C. Zanni-Merk, R. Howlett & L. Jain (Eds.), *Procedia computer science* (pp. 241–250). Elsevier B.V. https://doi.org/10.1016/j.procs.2021.08.025

Beneficence. (n.d.). Merriam-Webster. Retrieved December 10, 2022, from https://www.merriam-webster.com/dictionary/beneficence

Benn, S., Abratt, R., & O'Leary, B. (2016). Defining and identifying stakeholders: Views from management and stakeholders. *South African Journal of Business Management*, *47*(2), 1–11. https://doi.org/10.4102/sajbm.v47i2.55

BMW. (2020). *The path to autonomous driving*. Retrieved December 12, 2022, from https://www.bmw.com/en/automotive-life/autonomous-driving.html

Brennen, A. (2020). What do people really want when they say they want "explainable AI?" We asked 60 stakeholders. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3334480.3383047

Burget, M., Bardone, E., & Pedaste, M. (2017). Definitions and conceptual dimensions of Responsible Research and Innovation: A literature review. *Science and Engineering Ethics*, *23*(1). https://doi.org/10.1007/s11948-016-9782-1

Cetinic, E., & She, J. (2022). Understanding and creating art with AI: Review and outlook. *ACM Transactions on Multimedia Computing, Communications and Applications*, *18*(2). https://doi.org/10.1145/3475799

Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., & Malhotra, S. (2018, April 17). *Notes from the AI frontier: Applications and value of deep learning*. Retrieved December 12, 2022, from https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning

Chung, L., & Do Prado Leite, J. (2009). On non-functional requirements in software engineering (A. Borgida, V. Chaudhri, P. Giorgini & E. Yu, Eds.). *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *5600 LNCS*, 363–379. https://doi.org/10.1007/978-3-642-02463-4_19

Clinciu, M., & Hastie, H. (2019). A survey of explainable AI terminology. *NL4XAI 2019 - 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence, Proceedings of the Workshop*, 8–13.

Cooper, M. (2022). The impact of AI & how it is used in social media. Retrieved December 12, 2022, from https://statusbrew.com/insights/social-media-ai/#how-does-ai-work-in-social-media

Corbin, J., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/10.1007/BF00988593

Dalton-Brown, S. (2020). The ethics of medical AI and the physician-patient relationship. *Cambridge Quarterly of Healthcare Ethics*, *29*(1), 115–121. https://doi.org/10.1017/S0963180119000847

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved December 12, 2022, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Ding, P. (2019). Analysis of artificial intelligence (AI) application in sports. In J. Tseng, S. Cheung & I. Kotenko (Eds.), *Journal of physics: Conference series*. Institute of Physics Publishing. https://doi.org/10.1088/1742-6596/1302/3/032044

Downey, A. (2019). NHS bosses meet with tech giants to discuss commercial patient database. Retrieved December 8, 2022, from https://www.digitalhealth.net/2019/12/nhs-bosses-meet-with-tech-giants-to-discuss-commercial-patient-database/

Dybå, T., & Dingsøyr, T. (2008). Empirical studies of agile software development: A systematic review. *Information and Software Technology*, *50*(9-10), 833–859. https://doi.org/10.1016/j.infsof.2008.01.006

El-Haddadeh, R., Fadlalla, A., & Hindi, N. (2021). Is there a place for responsible artificial intelligence in pandemics? A tale of two countries. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10140-w

Ellahham, S. (2020). Artificial intelligence: The future for diabetes care. *The American Journal of Medicine*, *133*(8), 895–900. https://doi.org/https://doi.org/10.1016/j.amjmed.2020.03.033

European Commission. (2016, April 27). *Regulation (EU) 2016/679 of the European Parliament and of the Council*. Retrieved December 13, 2022, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

European Commission. (2021, April 21). Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Retrieved December 13, 2022, from https://artificialintelligenceact.eu/the-act/

Ford, R., & Richardson, W. (1994). Ethical decision making: A review of the empirical literature. *Journal of Business Ethics*, *13*(3), 205–221. https://doi.org/10.1007/BF02074820

Forman, C., & van Zeebroeck, N. (2019). Digital technology adoption and knowledge flows within firms: Can the Internet overcome geographic and technological distance? *Research Policy*, *48*(8), 103697. https://doi.org/https://doi.org/10.1016/j.respol.2018.10.021

Freeman, R., & Reed, D. (1983). Stockholders and stakeholders: A new perspective on corporate governance. *California Management Review*, *25*. https://doi.org/10.2307/41165018

Freeman, R. (1984). *Strategic management: A stakeholder approach*. Pitman.

Fruh, K., Archer, A., & Wojtowicz, J. (2022). Sportswashing: Complicity and corruption. *Sport, Ethics and Philosophy*, 1–18. https://doi.org/10.1080/17511321.2022.2107697

Gandjour, A., & Lauterbach, K. (2003). Utilitarian theories reconsidered: Common misconceptions, more recent developments, and health policy implications. *Health Care Analysis*, *11*(3), 229–244. https://doi.org/10.1023/B:HCAN.0000005495.81342.30

Gibbens, S. (2022). Is your favorite 'green' product as eco-friendly as it claims to be? *National Geographic*. Retrieved December 14, 2022, from https://www.nationalgeographic.com/environment/article/what-is-greenwashing-how-to-spot

Gillon, R. (1994). Medical ethics: Four principles plus attention to scope. *BMJ*, *309*(6948), 184. https://doi.org/10.1136/bmj.309.6948.184

Gillon, R. (2003). Ethics needs principles - four can encompass the rest - and respect for autonomy should be "first among equals". *Journal of Medical Ethics*, *29*(5), 307–312. https://doi.org/10.1136/jme.29.5.307

Glinz, M. (2007). On non-functional requirements. *15th IEEE International Requirements Engineering Conference (RE 2007)*, 21–26. https://doi.org/10.1109/RE.2007.45

Graham, N. (2020). Artificial intelligence: The next evolution of digital transformation. Retrieved December 15, 2022, from https://www.businessgoing.digital/artificial-intelligence-the-next-evolution-of-digital-transformation/

Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J., Yordanova, K., Vered, M., Nair, R., Abreu, P., Blanke, T., Pulignano, V., Prior, J., Lauwaert, L., Reijers, W., Depeursinge, A., Andrearczyk, V., & Müller, H. (2022). A global taxonomy of interpretable AI: Unifying the terminology for the technical and social sciences. *Artificial Intelligence Review*. https://doi.org/10.1007/s10462-022-10256-8

Greenhalgh, T., & Peacock, R. (2005). Effectiveness and efficiency of search methods in systematic reviews of complex evidence: Audit of primary sources. *British Medical Journal*, *331*(7524), 1064–1065. https://doi.org/10.1136/bmj.38636.593461.68

Grober, T., & Grober, O. (2020). Improving the efficiency of farm management using modern digital technologies. In R. D. & I. S. (Eds.), *E3s web of conferences*. EDP Sciences. https://doi.org/10.1051/e3sconf/202017513003

Hansen, S. (2010). Multicriteria mapping of stakeholder preferences in regulating nanotechnology. *Journal of Nanoparticle Research*, *12*(6), 1959–1970. https://doi.org/10.1007/s11051-010-0006-3

Hillman, A., & Keim, G. (2001). Shareholder value, stakeholder management, and social issues: What's the bottom line? *Strategic Management Journal*, *22*(2), 125–139. https://doi.org/10.1002/1097-0266(200101)22:2⟨125::AID-SMJ150⟩3.0.CO;2-H

Hoory, L., & Bottorff, C. (2022). What is a stakeholder analysis? Everything you need to know. *Forbes Advisor*. Retrieved December 15, 2022, from https://www.forbes.com/advisor/business/what-is-stakeholder-analysis/

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L., & Aerts, H. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5

Huang, J., & Fox, M. An ontology of trust - formal semantics and transitivity. In: 2006, 259–270. https://doi.org/10.1145/1151454.1151499.

IEEE. (2016). *Ethically aligned design*. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *IEEE Standards v1*.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, *2*(4), 230–243. https://doi.org/10.1136/svn-2017-000101

Justice. (n.d.). Merriam-Webster. Retrieved December 11, 2022, from https://www.merriam-webster.com/dictionary/justice

Kane, P. (2022). Yass marine! Here come rainbow bullets and brunches for Pride's annual pinkwashing. *The Guardian*. Retrieved December 14, 2022, from https://www.theguardian.com/world/2022/jun/07/yass-marine-here-come-rainbow-bullets-and-beers-for-prides-annual-pinkwashing

Keathley-Herring, H., Van Aken, E., Gonzalez-Aleu, F., Deschamps, F., Letens, G., & Orlandini, P. (2016). Assessing the maturity of a research area: Bibliometric review and proposed framework. *Scientometrics*, *109*(2), 927–951. https://doi.org/10.1007/s11192-016-2096-x

Kermany, D., Goldbaum, M., Cai, W., Valentim, C., Liang, H., Baxter, S., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., ... Zhang, K. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, *172*(5), 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Khatter, K., & Kalia, A. (2013). Impact of non-functional requirements on requirements evolution. *International Conference on Emerging Trends in Engineering and Technology, ICETET*, 61–68. https://doi.org/10.1109/ICETET.2013.15

Kretschmer, T., & Khashabi, P. (2020). Digital transformation and organization design: An integrated approach. *California Management Review*, *62*(4), 86–104. https://doi.org/10.1177/0008125620940296

Kumar, P., Kumar, R., Gupta, G., Tripathi, R., Jolfaei, A., & Najmul Islam, A. (2023). A blockchain-orchestrated deep learning approach for secure data transmission in IoT-enabled healthcare system. *Journal of Parallel and Distributed Computing*, *172*, 69–83. https://doi.org/10.1016/j.jpdc.2022.10.002

Kundisch, D., Muntermann, J., Oberländer, A., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2022). An update for taxonomy designers: Methodological guidance from information systems research. *Business and Information Systems Engineering*, *64*(4), 421–439. https://doi.org/10.1007/s12599-021-00723-x

Lang, G. (2004). A dilemma for objective act-utilitarianism. *Politics, Philosophy & Economics*, *3*(2), 221–239. https://doi.org/10.1177/1470594X04042966

Lardi, K. (2021). Understanding the value of artificial intelligence solutions in your business. *Forbes*. Retrieved December 12, 2022, from https://www.forbes.com/sites/forbesbusinesscouncil/2021/01/26/understanding-the-value-of-artificial-intelligence-solutions-in-your-business/?sh=75c7364657f6

Lasi, H., Fettke, P., Kemper, H., Feld, T., & Hoffmann, M. (2014). Industry 4.0. *Business and Information Systems Engineering*, *6*(4), 239–242. https://doi.org/10.1007/s12599-014-0334-4

Lu, Y. (2017). Industry 4.0: A survey on technologies, applications and open research issues. *Journal of Industrial Information Integration*, *6*, 1–10. https://doi.org/10.1016/j.jii.2017.04.005

Marshall, A., & Goehring, B. (2020, November 20). *The business value of AI: Peak performance during the pandemic*. Retrieved December 12, 2022, from https://www.ibm.com/thought-leadership/institute-business-value/report/ai-value-pandemic

Martinez, A., Phillips, S., Carrilho, E., Thomas III, S., Sindi, H., & Whitesides, G. (2008). Simple telemedicine for developing regions: Camera phones and paper-based microfluidic devices for real-time, off-site diagnosis. *Analytical Chemistry*, *80*(10), 3699–3707. https://doi.org/10.1021/ac800112r

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development?, 729–733. https://doi.org/10.1145/3236024.3264833

Mikalef, P., & Gupta, M. (2021). Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, *58*(3), 103434. https://doi.org/10.1016/j.im.2021.103434

Miller, C., & Coldicutt, R. (2019). People, power and technology: The tech workers' view. https://doteveryone.org.uk/report/workersview

Mitchell, R., Agle, B., & Wood, D. (1997). Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *Academy of Management Review*, *22*(4), 853–886. https://doi.org/10.5465/AMR.1997.9711022105

Mohapatra, B., Tripathy, S., Singhal, D., & Saha, R. (2022). Significance of digital technology in manufacturing sectors: Examination of key factors during Covid-19. *Research in Transportation Economics*, *93*. https://doi.org/10.1016/j.retrec.2021.101134

Moor, J. (2006). The Dartmouth College Artificial Intelligence Conference: The next fifty years. *AI Magazine*, *27*(4), 87–91.

Myers, A. (2022). How AI is making autonomous vehicles safer. *Stanford University Human-Centered Artificial Intelligence – News*. https://hai.stanford.edu/news/how-ai-making-autonomous-vehicles-safer

Netflix Research. (n.d.). Machine learning. Retrieved December 12, 2022, from https://research.netflix.com/research-area/machine-learning

Norwegian Register. (n.d.). *About the Register for Scientific Journals, Series and Publishers*. Retrieved December 9, 2022, from https://kanalregister.hkdir.no/publiseringskanaler/Om

O'Mara, B., Gill, G., Babacan, H., & Donahoo, D. (2012). Digital technology, diabetes and culturally and linguistically diverse communities: A case study with elderly women from the Vietnamese community. *Health Education Journal*, *71*(4), 491–504. https://doi.org/10.1177/0017896911407054

Page, M., McKenzie, J., Bossuyt, P., Boutron, I., Hoffmann, T., Mulrow, C., Shamseer, L., Tetzlaff, J., Akl, E., Brennan, S., Chou, R., Glanville, J., Grimshaw, J., Hróbjartsson, A., Lalu, M., Li, T., Loder, E., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*. https://doi.org/10.1136/bmj.n71

Papagiannidis, E., Enholm, I., Dremel, C., Mikalef, P., & Krogstie, J. (2022). Toward AI governance: Identifying best practices and potential barriers and outcomes. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-022-10251-y

Personopplysningsloven. (2018, July 20). Lov om behandling av personopplysninger. Retrieved December 13, 2022, from https://lovdata.no/dokument/NL/lov/2018-06-15-38

Privacy. (n.d.). Merriam-Webster. Retrieved December 12, 2022, from https://www.merriam-webster.com/dictionary/privacy

PwC. (2017). *Sizing the prize: What's the real value of AI for your business and how can you capitalise?* Retrieved December 12, 2022, from https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html

Quest, H., Cauz, M., Heymann, F., Rod, C., Perret, L., Ballif, C., Virtuani, A., & Wyrsch, N. (2022). A 3D indicator for guiding AI applications in the energy sector. *Energy and AI*, *9*. https://doi.org/10.1016/j.egyai.2022.100167

Raghavan, P. (2020). How AI is powering a more helpful Google. Retrieved December 12, 2022, from https://blog.google/products/search/search-on/

Rajkomar, A., Oren, E., Chen, K., Dai, A., Hajaj, N., Hardt, M., Liu, P., Liu, X., Marcus, J., Sun, M., Sundberg, P., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G., Irvine, J., Le, Q., Litsch, K., ... Dean, J. (2018). Scalable and

accurate deep learning with electronic health records. *npj Digital Medicine*, *1*(1). https://doi.org/10.1038/s41746-018-0029-1

Ransbotham, S., Kiron, D., Candelon, F., Khodabandeh, S., Chu, M., & LaFountain, B. (2022, November 1). *AI empowers employees, not just companies*. Retrieved December 12, 2022, from https://www.bcg.com/publications/2022/the-value-of-ai-for-individuals

Reed, M., Graves, A., Dandy, N., Posthumus, H., Hubacek, K., Morris, J., Prell, C., Quinn, C., & Stringer, L. (2009). Who's in and why? A typology of stakeholder analysis methods for natural resource management. *Journal of Environmental Management*, *90*(5), 1933–1949. https://doi.org/10.1016/j.jenvman.2009.01.001

Reinoso, S. (2021). *Auditors create trust*. Retrieved December 13, 2022, from https://dobetter.esade.edu/en/auditors-create-trust

Responsible. (n.d.). Cambridge Dictionary. Retrieved November 29, 2022, from https://dictionary.cambridge.org/dictionary/english/responsible

Ross, J., Mocker, M., & Beath, C. (2018, June 21). Five building blocks of digital transformation. Retrieved December 15, 2022, from https://cisr.mit.edu/publication/2018_0601_BuildingBlocks_RossMockerBeath

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

Samuel, A. (1960). Some moral and technical consequences of automation: A refutation. *Science*, *132*(3429), 741–742. https://doi.org/10.1126/science.132.3429.741

Sarker, I. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3). https://doi.org/10.1007/s42979-021-00592-x

Schomberg, R. (2013). A vision of Responsible Research and Innovation. In R. Owen, J. Bessant & M. Heintz (Eds.), *Responsible Innovation: Managing the responsible emergence of science and innovation in society* (pp. 51–74). https://doi.org/10.1002/9781118551424.ch3

Seppälä, S., Schreiber, Y., & Ruttenberg, A. (2014). Textual and logical definitions in ontologies. In D. Malone, A. Ruttenberg, M. Haendel, M. Brochhausen, R. Boyce, W. Hogan, C. Stoeckert, J. Zheng, P. Empey, P. Ray, S. Seppälä & M. Brochhausen (Eds.), *CEUR workshop proceedings* (pp. 35–41). CEUR-WS.

Shubhendu, S., & Vijay, J. (2013). Applicability of artificial intelligence in different fields of life. *International Journal of Scientific Engineering and Research*.

Singapore Personal Data Protection Commission. (2020). *Model artificial intelligence governance framework*. Retrieved November 29, 2022, from https://

www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

Singh, P., & Sharma, A. (2023). Using AI-based approaches in health care for predicting health issues in pregnant women (A. Khanna, D. Gupta, V. Kansal, G. Fortino & A. Hassanien, Eds.). *Lecture Notes in Networks and Systems*, *479*, 283–295. https://doi.org/10.1007/978-981-19-3148-2_24

Son, D., Lee, J., Qiao, S., Ghaffari, R., Kim, J., Lee, J., Song, C., Kim, S., Lee, D., Jun, S., Yang, S., Park, M., Shin, J., Do, K., Lee, M., Kang, K., Hwang, C., Lu, N., Hyeon, T., & Kim, D. (2014). Multifunctional wearable devices for diagnosis and therapy of movement disorders. *Nature Nanotechnology*, *9*(5), 397–404. https://doi.org/10.1038/nnano.2014.38

Taddy, M. (2018). *The technological elements of artificial intelligence* (Working Paper No. 24301). National Bureau of Economic Research. https://doi.org/10.3386/w24301

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, *14*(3), 207–222. https://doi.org/https://doi.org/10.1111/1467-8551.00375

Trocin, C., Mikalef, P., Papamitsiou, Z., & Conboy, K. (2021). Responsible AI for digital health: A synthesis and a research agenda. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10146-4

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105–112. https://doi.org/10.1007/s10676-009-9187-9

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. *Knowledge Engineering Review*, *11*(2), 93–136. https://doi.org/10.1017/s0269888900007797

Vaishya, R., Javaid, M., Khan, I., & Haleem, A. (2020). Artificial intelligence (AI) applications for COVID-19 pandemic. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, *14*(4), 337–339. https://doi.org/10.1016/j.dsx.2020.04.012

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, *11*(1), 233. https://doi.org/10.1038/s41467-019-14108-y

Wang, J., Ma, X., Zhang, J., & Zhao, X. (2022). Impacts of digital technology on energy sustainability: China case study. *Applied Energy*, *323*. https://doi.org/10.1016/j.apenergy.2022.119329

Wang, W. (2020). Research on design and innovation based on network platform and digital technology. *Proceedings - 2020 International Conference on Innovation Design and Digital Technology, ICIDDT 2020*, 401–404. https://doi.org/10.1109/ICIDDT52279.2020.00080

Ward, A. (2012). Scientists probe human nature–and discover we are good, after all. *Scientific American*. Retrieved December 13, 2022, from https://www.scientificamerican.com/article/scientists-probe-human-nature-and-discover-we-are-good-after-all/

Wearing, D. (2022). A game of two halves: How 'sportswashing' benefits Qatar and the west. *The Guardian*. Retrieved December 14, 2022, from https://www.theguardian.com/commentisfree/2022/nov/16/sportswashing-qatar-west-world-cup-regime

Weisha, Z. (2021). The impact of digital technology on enterprise innovation by knowledge management perspective. *Proceedings - 2021 International Conference on Big Data and Intelligent Decision Making, BDIDM 2021*, 29–32. https://doi.org/10.1109/BDIDM53834.2021.00013

Werker, C. (2020). Assessing Responsible Research and Innovation (RRI) systems in the digital age. In E. Yaghmaei & I. van de Poel (Eds.), *Assessment of Responsible Innovation: Methods and practices* (1st ed.). Routledge.

Widén-Wulff, G., & Ginman, M. (2004). Explaining knowledge sharing in organizations through the dimensions of social capital. *Journal of Information Science*, *30*(5), 448–458. https://doi.org/10.1177/0165551504046997

Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, *131*(3410), 1355–1358. https://doi.org/10.1126/science.131.3410.1355

Wiesing, U. (2020). The hippocratic oath and the declaration of geneva: Legitimisation attempts of professional conduct. *Medicine, Health Care and Philosophy*, *23*(1), 81–86. https://doi.org/10.1007/s11019-019-09910-w

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/2601248.2601268

Wolford, B. (n.d.). *What are the GDPR fines?* Retrieved December 13, 2022, from https://gdpr.eu/fines/

Xu, L., Xu, E., & Li, L. (2018). Industry 4.0: State of the art and future trends. *International Journal of Production Research*, *56*(8), 2941–2962. https://doi.org/10.1080/00207543.2018.1444806

Yudha, S., Tjahjono, B., & Kolios, A. (2018). A PESTLE policy mapping and stakeholder analysis of Indonesia's fossil fuel energy industry. *Energies*, *11*(5). https://doi.org/10.3390/en11051272

Zaheer, N., & Trkman, P. (2017). An information sharing theory perspective on willingness to share information in supply chains. *The International Journal of Logistics Management*, *28*(2), 417–443. https://doi.org/10.1108/IJLM-09-2015-0158

Appendix A Search terms

Search terms used for the two database searches.

A.1 Scoping study

Scopus

```
ALL ( "responsible AI" )
OR (
TITLE-ABS-KEY ( "machine learning" OR "artificial intelligence" OR
"ai" )
AND TITLE-ABS-KEY ( "responsible" )
)
AND (
TITLE-ABS-KEY ( business* )
)
AND (
LIMIT-TO ( DOCTYPE , "ar" )
OR LIMIT-TO ( DOCTYPE , "cp" )
OR LIMIT-TO ( DOCTYPE , "ch" )
OR LIMIT-TO ( DOCTYPE , "ed" )
)
AND (
LIMIT-TO ( LANGUAGE , "English" )
)
```

Web of Science

```
(ALL=("responsible AI") OR (TS=("machine learning" OR "artificial
intelligence" OR ai) AND TS=(responsible))) AND TS=(business*) AND
LA=(English)
```

A.2 Main study

Scopus

```
TITLE-ABS-KEY ( "responsible AI" OR "responsible artificial
intelligence" )
AND (
LIMIT-TO ( DOCTYPE , "ar" )
OR LIMIT-TO ( DOCTYPE , "cp" )
OR LIMIT-TO ( DOCTYPE , "ch" )
OR LIMIT-TO ( DOCTYPE , "ed" )
)
AND (
LIMIT-TO ( LANGUAGE , "English" )
)
```

Web of Science

```
(ALL=("responsible AI") OR ALL=("responsible artificial
intelligence")) AND LA=(English) AND DT=(Article OR Book Chapter OR
Editorial Material OR Proceedings Paper OR Review)
```

Appendix B Data extraction example

Table 10: Example of data extraction applied to Liu et al. (2021). For a description of each column, see Table 3.

Column	**Data**
Metadata	
Author	Liu R., Gupta S., Patel P.
Title	The Application of the Principles of Responsible AI on Social Media Marketing for Digital Health
Source	Information Systems Frontiers
Source rank	1
Year	2021
Abstract	Social media enables medical professionals and authorities to share, disseminate, monitor, and manage health-related information digitally through online communities such as Twitter and Facebook. Simultaneously, artificial intelligence (AI) powered social media offers digital capabilities for organizations to select, screen, detect and predict problems with possible solutions through digital health data. Both the patients and healthcare professionals have benefited from such improvements. However, arising ethical concerns related to the use of AI raised by stakeholders need scrutiny which could help organizations obtain trust, minimize privacy invasion, and eventually facilitate the responsible success of AI-enabled social media operations. This paper examines the impact of responsible AI on businesses using insights from analysis of 25 in-depth interviews of health care professionals. The exploratory analysis conducted revealed that abiding by the responsible AI principles can allow healthcare businesses to better take advantage of the improved effectiveness of their social media marketing initiatives with their users. The analysis is further used to offer research propositions and conclusions, and the contributions and limitations of the study have been discussed. © 2021, The Author(s).
Methodology	Qualitative interviews Conceptual framework
Sample	25 health care professionals (healthcare managers and employees) in China

Continued on next page...

Column	Data
Context	China, healthcare
Contribution	Found that following responsible AI principles improves the effectiveness of social media marketing initiatives of healthcare businesses
Key concept definitions	
AI	"A broad range of techniques and approaches of computer science to simulate human intelligence in machines that are programmed for thinking like humans and mimic human behaviours, capable of performing tasks" (p.3-4)
Responsible AI	"Seeks to support the design, implementation, and use of ethical, transparent, and accountable AI solutions, aiming at reducing biases, facilitating fairness, equality, interpretability, and explainability (Trocin et al., 2021)" (p.3) "Eitel-Porter (2021) defined the term as the practice of using AI with good intention, to empower employees and businesses and create a fair environment for customers and society, ultimately enabling organizations to generate trust and bring AI deployments to scale" (p.3) "Taylor et al. (2018) regarded responsible AI as an umbrella term, which investigates legal, ethical, and moral viewpoints of autonomous algorithms or applications of AI that may be crucial to safety or may impact people's lives in a disruptive way" (p.3)
AI ethics	*
Synonyms	"Responsible AI has been used interchangeably with ethical AI or the responsible use of AI." (p.3)
AI system	*
Research questions	
Principles	From combining Clarke (2019), Microsoft AI (2020) and Floridi et al. (2018): - Fairness

Continued on next page...

Column	Data
	"Should not lead to discriminatory influences on humans associated with race, ethnic origin, religion, gender, sexual orientation, disability or other situations (Benjamins et al., 2019)" (p.14) - Inclusiveness - Reliability and safety - Transparency - Privacy and security - Beneficence - Non-maleficence - Autonomy
Antecedents	*
Outcomes	Transparency: - "Maintain the citizen's rights to be informed" (p.21) Fairness: - "Increase users' willingness to engage in the technology" (p.20) Reliability and safety: - "Users are more likely to perceive the usefulness of the technology to be willing to participate in it." (p.21) Beneficence: - "If users' realization of the usefulness of responsible AI for health management increased, the acceptance of the technology would soon improve" (p.21) Autonomy: - Increase "users' acceptability and intention to use the technology" (p.21)

Continued on next page...

Column	Data
	"How well ethics are treated will ultimately decide how much people will embrace technology in the future" (p.21)
Business advantages	"Companies might overlook the financial rewards of responsible AI and treat it as the sole way to avoid risks (Cheng et al., 2021)." (p.2) Privacy and security: - "A high level of good credibility and corporate reputation" (p.21) Helps to "sustain a long-term relationship consumers" (p.22) "It is believed that investing in social values based on trust, mutuality, and respect could enable long-term organizational benefits such as corporate well-being and innovativeness (Widén-Wulff & Ginman, 2004)." (p.22)
Barriers	"The attempts of these principles to translate into AI practices is still in their infancy (Benjamins, 2020; Cheng et al., 2021; Scantamburlo et al., 2020)" (p.2) "It seems to be difficult to apply [responsible AI] to the practical hierarchy or to apply it to industries broadly." (p.4) "Most responsible AI guidelines are useful and can help shape policy but are not easily enforceable" (p.4)
Facilitators	*
Enabling	Fairness: - "Pay close attention to less privileged groups" (p.20) Inclusiveness: - "Involve more diverse audiences, respect their distinctive characteristics, and encourage active participation." (p.20) Transparency: - Make AI systems "relatively transparent" (p.20)

Continued on next page...

Column	Data
	Autonomy: - Different people (both across ages and the same age) may have different limits for how much assistance they want from an AI before they feel like they are losing autonomy - AI systems must therefore adapt to each person Privacy and security: - "Seek strict privacy protection" (p.21) To sustain long-term relationships, a company should "concentrate on building trust, collecting, and ethically processing the data" (p.22)
Added after initial test	
Cases	"An artificially intelligent chatbot on a mobile app was created to give British people diagnostic advice on common ailments without human interaction (Olson, 2018)" (p.2)
Other relevant data	"Many common pitfalls raise risks for an organization: rushed development, a lack of technical understanding, improper quality assurance, use of AI outside the original context, improper blends of data, and reluctance by employees to raise concerns (Eitel-Porter, 2021)" (p.4) "Existing approaches to the implementation of ethical reasoning can be divided into three main categories: - Top-down approaches are to implement a given ethical theory and apply it to a particular case; - bottom-up approaches are to aggregate sufficient observations of similar situations into a decision and infer general rule from these independent cases; - hybrid approaches combine the benefits of bottom-up and top-down approaches in support of a careful moral reflection (Singer, 2020)" (p.4)

Continued on next page...

Column	Data
	"Microsoft has launched the Office of Responsible AI (ORA) and the AI, Ethics, and Effects in Engineering and Research (Aether) Committee to put responsible AI principles into practice (Microsoft AI, 2020). " (p.5) "The technology acceptance model (TAM) [...] assumes that an individuals' technology acceptance is determined by two major factors: perceived usefulness and perceived ease of use (King & He, 2006)." (p.11) "Trust is the most crucial antecedent and motivation for consumers' willingness to share (Zaheer & Trkman, 2017)." (p.22)
Own summary	
Summary	Combines three sources of principles to create their own set, and contains significant data / opinions on what the principles (and responsible AI in general) can lead to (mostly - increased trust).

* No data in this paper.

Appendix C Overview of reviewed papers

Table 11: Overview of the reviewed papers

Article	Summary	Context	Methodology
Anagnostou et al. (2022)	Looks at Business management, Transportation, Healthcare, E-Government & public sector and Information technology, to see what issues these industries face regarding responsible AI, and what principles are needed to handle these issues.	Global, multiple sectors	Literature review
Balagué (2021)	Looks at negative impacts of AI and how they can be managed	N/A*	Book‡
Barredo Arrieta et al. (2020)	Provides an in-depth, technical overview of explainable AI.	N/A*	Literature review
Bélisle-Pipon et al. (2022)	Reviews existing guidelines for responsible AI, focusing on methodology and stakeholder engagement.	N/A*	Review of guidelines
Benjamins et al. (2019)	Presents a set of principles for responsible AI, and a methodology for implementing them.	Large, multinational organisation	Case study
Borda et al. (2022)	Looks at how principles for responsible AI can have a positive effect on studies of diseases.	N/A*	Unknown†
Brand (2022)	Presents a framework for implementing responsible AI in government.	South Africa, government	Review of guidelines
Buhmann and Fieseler (2021)	Presents a high-level framework for implementing responsible AI	N/A*	Conceptual
Canca (2020)	Reviews existing principles for responsible AI, separating them into "core" and "instrumental" principles.	N/A*	Viewpoint
Chen (2020)	Presents a framework for implementing responsible AI.	N/A*	Conceptual
Cheng et al. (2021)	Presents a framework for implementing responsible AI.	N/A*	Conceptual
Clarke (2019)	Reviews existing guidelines for responsible AI, combining them to create 50 principles divided into 10 themes.	N/A*	Review of guidelines
Dignum (2017)	Presents the ART-principles for responsible AI.	Europe, labor	Expert discussion
Dignum (2019)	Presents a framework for connecting abstract principles to concrete implementations.	N/A*	Book‡
Dignum (2021)	Looks at how AI can be used responsibly in education.	Education	Unknown†
Doorn (2021)	Reviews how responsible AI can be used in the water domain.	Global, water domain	Literature review
Eitel-Porter (2021)	Presents a DevOps-look at how responsible AI can be implemented and handled in production systems.	N/A*	Viewpoint
Fjeld et al. (2020)	Provides detailed descriptions of existing principles for responsible AI.	N/A*	Review of guidelines

Continued on next page...

Article	Summary	Context	Methodology
Floridi and Cowls (2019)	Reviews existing principles for responsible AI, finding a convergence upon a set of common principles.	Global	Review of guidelines
Floridi et al. (2018)	Presents results from the AI4People initiative, including 20 action points for implementing responsible AI.	N/A*	Expert discussion
Gianni et al. (2022)	Presents a philosophical view of responsible AI, looking at how it can be achieved in a democratic way.	Global	Review of AI strategies
Gupta et al. (2021)	Looks at AI in healthcare through a perceived risk theory-lens.	India, healthcare	Quantitative survey
Hacker and Passoth (2022)	Looks at current and proposed laws regulating AI in Europe.	Europe, laws and regulation	Review of laws
Hagendorff (2020)	Draws a connection between how common a principle is, and how technical its solution is.	N/A*	Review of guidelines
Havrda and Rakova (2020)	Presents an impact assessment framework to operationalize principles for responsible AI.	N/A*	Conceptual
Henriksen et al. (2021)	Looks at how developers consider different accountability measures.	Scandinavia, AI developers	Case study
Jakesch et al. (2022)	Finds that stakeholders and AI developers have differing views on the importance of AI principles.	The US	Quantitative survey
Jobin et al. (2019)	Reviews existing guidelines for responsible AI, using that to create a new set of principles.	Western countries	Review of guidelines
Kumar et al. (2021)	Looks at connections between the responsibility of AI and the perceived and actual value of a solution.	India, healthcare	Qualitative interviews, quantitative survey
Liu et al. (2021)	Reviews existing guidelines for responsible AI, using that to create a new set of principles.	China, healthcare	Qualitative interviews
Q. Lu et al. (2022)	Presents a detailed list of methods for implementing responsible AI.	N/A*	Conceptual
Lukkien et al. (2021)	Reviews existing AI principles within the field of long-term healthcare.	Healthcare	Literature review
Merhi (2022)	Looks at existing barriers preventing AI from being responsible.	N/A*	Literature review, qualitative interviews
Mikalef et al. (2022)	Presents the authors' view on the direction of future AI research.	N/A*	Viewpoint
Minkkinen et al. (2021)	Reviews AI strategies in place in Europe, and how that impacts the AI ecosystem.	Europe, AI strategies	Review of AI strategies
Morley et al. (2020)	Presents a detailed list of tools for integrating ethics in AI systems.	N/A*	Literature review
Morley et al. (2021)	Looks at barriers preventing AI from being responsible, as well as ways to overcome them.	The UK	Qualitative interviews, quantitative survey
Nauck (2019)	Presents recommendations for managers interested in making AI responsible.	N/A*	Viewpoint
Nevanperä et al. (2021)	Reviews existing guidelines.	N/A*	Review of guidelines
Papagiannidis, Mikalef et al. (2022)	Looks at business advantages that can be gained from making AI responsible.	Nordic countries	Quantitative survey

Continued on next page...

Article	Summary	Context	Methodology
Peters et al. (2020)	Presents two frameworks for implementing responsible AI, and applies them to a case study.	Global, healthcare	Case study, qualitative interviews
Rakova et al. (2021)	Presents a detailed list of organisational changes that enable responsible AI.	Global, AI developers	Qualitative interviews, expert workshop
Rizinski et al. (2022)	Maps existing ethical challenges of finance to AI ethics.	Global, finance	Qualitative interviews
Rothenberger et al. (2019)	Presents a set of principles for responsible AI, and uses experts and laypersons to prioritize them.	N/A*	Literature review, qualitative interviews, quantitative survey
Ryan and Stahl (2021)	Presents a detailed list of normative recommendations for creating responsible AI.	N/A*	Review of guidelines
Siala and Wang (2022)	Summarizes current work on responsible AI within the field of healthcare.	Global, healthcare	Literature review
Thelisson (2018)	Presents ethical theories and principles for responsible AI.	N/A*	Unknown†
Vakkuri et al. (2022)	Compares actual implementations of responsible AI to theoretical guidelines.	Finland, AI developers	Qualitative interviews
van Bruxvoort and van Keulen (2021)	Presents a framework for implementing responsible AI.	The Netherlands, government	Case study
Vetrò et al. (2019)	Presents a set of principles for developing responsible AI agents.	N/A*	Unknown†
W. Wang et al. (2021)	Looks at how responsible AI impacts technology adoption.	China, healthcare	Quantitative survey
Y. Wang et al. (2020)	Draws a connection between responsible AI and corporate social responsibility.	Western countries	Literature review
Werder et al. (2022)	Looks at ways to responsibly handle data, to use for creating responsible AI.	N/A*	Conceptual
Wright and Schultz (2018)	Looks at how automating business processes impact stakeholders.	N/A*	Conceptual

* Not applicable to this paper.
† No methodology provided.
‡ No methodology applicable.

Appendix D Principles of reviewed papers

Table 12: Core principles used by the reviewed papers.

Article	Based on	Autonomy	Beneficence	Non-maleficence	Justice
Anagnostou et al. (2022)	–	–	–	Security	Fairness, Empathy
Balagué (2021)	Beauchamp and Childress (2001)	Autonomy	Beneficence	Non-maleficence	Justice
Barredo Arrieta et al. (2020)	–	–	Human-centeredness	–	Fairness
Benjamins et al. (2019)	–	–	Human-centric	Privacy and Security by Design	Fair
Borda et al. (2022)	Floridi et al. (2018)	Autonomy	Beneficence	Non-maleficence	Justice
Brand (2022)	–	–	–	–	Fairness, Respect for human rights
Buhmann and Fieseler (2021)	–	–	Do good	Avoid harm	–
Canca (2020)	–	Respect for autonomy	Beneficence	–	Justice
Cheng et al. (2021)	–	–	–	Reliability and Safety, Privacy and Security	Fairness, Inclusiveness
Clarke (2019)	–	"Ensure Human Control"	"Assess Positive and Negative Impacts and Implications", "Complement Humans"	"Assess Positive and Negative Impacts and Implications", "Ensure Human Safety and Wellbeing", "Embed Quality Assurance", "Exhibit Robustness and Resilience"	"Ensure Consistency with Human Values and Human Rights"
Dignum (2017)	–	–	–	–	–
Dignum (2019)	Dignum (2017)	–	–	–	–
Dignum (2021)	Dignum (2017)	–	–	–	–
Doorn (2021)	–	–	–	Non-maleficence	Justice and fairness
Eitel-Porter (2021)	–	–	–	–	Fairness
Fjeld et al. (2020)	–	Human control of technology	–	Safety and security	Fairness and non-discrimination
Floridi et al. (2018)	–	Autonomy	Beneficence	Non-maleficence	Justice

Continued on next page...

Article	Based on	Autonomy	Beneficence	Non-maleficence	Justice
Floridi and Cowls (2019)	–	Autonomy	Beneficence	Non-maleficence	Justice
Gianni et al. (2022)	–	Controllability	–	Safety, Security	–
Gupta et al. (2021)	–	–	–	Privacy and Security	Equality
Hacker and Passoth (2022)	–	–	–	–	Connections to algorithmic fairness
Hagendorff (2020)	–	–	–	–	Fairness
Havrda and Rakova (2020)	Fjeld et al. (2020)	Human control of technology, Promotion of Human Values	Promotion of human values	Safety and security	Fairness and Non-discrimination, Promotion of Human Values
Henriksen et al. (2021)	Jobin et al. (2019)	–	–	Non-maleficence	Justice and fairness
Jakesch et al. (2022)	Jobin et al. (2019)	Freedom and human autonomy	Social good (Beneficence)*	Safety (Non-maleficence)*, Dignity	Justice and fairness
Jobin et al. (2019)	–	Freedom and autonomy	Beneficence	Non-maleficence, Dignity	Justice and fairness, Solidarity
Kumar et al. (2021)	Dignum (2017)	–	–	–	–
Liu et al. (2021)	Clarke (2019), Microsoft AI and Floridi et al. (2018)	Autonomy	Beneficence	Reliability and safety, Non-maleficence	Fairness, Inclusiveness
Q. Lu et al. (2022)	–	Human Control of Technology, Promotion of Human Values	Promotion of Human Values	Safety and Security	Fairness and Non-discrimination, Promotion of Human Values
Lukkien et al. (2021)	–	Autonomy, Informed consent	–	–	Justice, Fairness
Merhi (2022)	Mikalef et al. (2022)	Human oversight	Societal and environmental well-being	Robustness and safety, Data governance	Fairness
Mikalef et al. (2022)	–	Human oversight	Societal and environmental well-being	Robustness and safety, Data governance	Fairness
Morley et al. (2020)	Floridi et al. (2018)	Autonomy	Beneficence	Non-maleficence	Justice
Morley et al. (2021)	Floridi et al. (2018)	Autonomy	Beneficence	Non-maleficence	Justice

Continued on next page...

Article	Based on	Autonomy	Beneficence	Non-maleficence	Justice
Nauck (2019)	–	–	"The benefit of society"	"Do no harm"	Fair, "AI should actively guard against bias or any other form of discrimination"
Papagiannidis, Mikalef et al. (2022)	–	–	Societal well-being, Human-centric AI	Robustness and safety, Data governance	Fairness
Peters et al. (2020)	Floridi et al. (2018)	Respect for autonomy	Beneficence	Non-maleficence	Justice
Rizinski et al. (2022)	OECD	–	"Inclusive growth, sustainable development and well-being"	Robustness, security and safety	"Human-centred values and fairness"
Rothenberger et al. (2019)	–	"An AI should have a purpose"	–	Robustness	"Bias should be minimized"
Ryan and Stahl (2021)	EU's HLEG and Jobin et al. (2019)	Freedom and autonomy	Beneficence	Non-maleficence, Dignity	Justice and fairness
Siala and Wang (2022)	–	–	Human centeredness	–	Inclusiveness, Fairness
Thelisson (2018)	–	"Meaningful human control"	"Use for socially beneficial purposes"	–	"Enhance fairness, freedom, fraternity and equality", "Avoid discrimination"
Vakkuri et al. (2022)	–	–	–	–	–
van Bruxvoort and van Keulen (2021)	Floridi et al. (2018)	Autonomy	Beneficence	Non-maleficence	Justice
Vetrò et al. (2019)	–	–	–	–	–
W. Wang et al. (2021)	Floridi et al. (2018)	Autonomy	Beneficence	Non-maleficence	Justice
Y. Wang et al. (2020)	–	–	–	Avoid ethical challenges	–
Werder et al. (2022)	–	–	–	–	Fairness
Papers including principle		*24*	*27*	*32*	*39*
Percentage of papers with principles		*50%*	*56%*	*67%*	*81%*
Percentage of all papers		*44%*	*50%*	*59%*	*72%*

* Parentheses show principles before pilot study.
Only papers presenting a set of principles are included in this table.

Table 13: Instrumental principles used by the reviewed papers.

Article	Transparency	Accountability	Trust	Sustainability	Privacy	Others
Anagnostou et al. (2022)	Transparency	Accountability	–	–	–	–
Balagué (2021)	–	–	–	–	–	–
Barredo Arrieta et al. (2020)	Explainability	Accountability	–	–	Privacy	–
Benjamins et al. (2019)	Transparent and explainable	–	–	–	Privacy and Security by Design	–
Borda et al. (2022)	Explicability	–	–	–	–	–
Brand (2022)	Transparency	Accountability	–	–	Privacy	–
Buhmann and Fieseler (2021)	–	–	–	–	–	–
Canca (2020)	–	–	–	–	–	–
Cheng et al. (2021)	Transparency	Accountability	–	–	Privacy and Security	–
Clarke (2019)	"Deliver Transparency and Auditability"	"Enforce, and Accept Enforcement of, Liabilities and Sanctions", "Ensure Accountability for Obligations"	–	–	–	–
Dignum (2017)	Transparency	Accountability, Responsibility	–	–	–	–
Dignum (2019)	Transparency	Accountability, Responsibility	–	–	–	–
Dignum (2021)	Transparency	Accountability, Responsibility	–	–	–	–
Doorn (2021)	Transparency	Responsibility and accountability	–	–	Privacy	–
Eitel-Porter (2021)	Transparency, Explainability	Accountability	–	–	Privacy	–
Fjeld et al. (2020)	Transparency and explainability	Accountability, Professional responsibility	–	–	Privacy	–

Continued on next page...

Article	Transparency	Accountability	Trust	Sustainability	Privacy	Others
Floridi et al. (2018)	Explicability	–	–	–	–	–
Floridi and Cowls (2019)	Explicability	–	–	–	–	–
Gianni et al. (2022)	Transparency	Accountability, Responsibility	Trust-worthiness	–	Privacy	–
Gupta et al. (2021)	Explainability	Answerability	–	–	Privacy and Security	–
Hacker and Passoth (2022)	Actionable explanations	–	–	–	–	Quality bench-marking, Interventions and co-design
Hagendorff (2020)	–	Accountability	–	–	Privacy	–
Havrda and Rakova (2020)	Transparency and explainability	Accountability, Professional responsibility	–	–	Privacy	–
Henriksen et al. (2021)	Transparency	Responsibility	–	–	Privacy	–
Jakesch et al. (2022)	Transparency	Accountability	(Trust)*	–	Privacy	–
Jobin et al. (2019)	Transparency	Responsibility	Trust	Sustainability	Privacy	–
Kumar et al. (2021)	Transparency	Accountability, Responsibility	–	–	–	–
Liu et al. (2021)	Transparency	–	–	–	Privacy and security	–
Q. Lu et al. (2022)	Transparency and Explainability	Accountability	–	–	Privacy	–
Lukkien et al. (2021)	Transparency	–	Trust	–	Privacy	–
Merhi (2022)	Transparency	Accountability	–	Societal and environmental well-being	–	Laws and regulations
Mikalef et al. (2022)	Transparency	Accountability	–	Societal and environmental well-being	–	Laws and regulations
Morley et al. (2020)	Explicability	–	–	–	–	–
Morley et al. (2021)	Explicability	–	–	–	–	–

Continued on next page...

Article	Transparency	Accountability	Trust	Sustainability	Privacy	Others
Nauck (2019)	Transparent, Decisions should be explained where possible	–	–	–	–	–
Papagiannidis, Mikalef et al. (2022)	Transparency	Accountability	–	Environmental	–	–
Peters et al. (2020)	Explicability	–	–	–	–	–
Rizinski et al. (2022)	Transparency and explainability	Accountability	–	"Inclusive growth, sustainable development and well-being"	–	–
Rothenberger et al. (2019)	Transparency	Responsibility	–	–	Protection of Data Privacy	–
Ryan and Stahl (2021)	Transparency	Responsibility	Trust	Sustainability	Privacy	–
Siala and Wang (2022)	Transparency	–	–	Sustainability	–	–
Thelisson (2018)	"Transparent and open manner"	–	–	–	–	–
Vakkuri et al. (2022)	Transparency	Accountability, Responsibility	–	–	–	Predictability
van Bruxvoort and van Keulen (2021)	Explicability	–	–	–	–	–
Vetrò et al. (2019)	Transparency and openness, Explainability and understandability	–	–	–	Privacy	–
W. Wang et al. (2021)	Explainability	–	–	–	–	–
Y. Wang et al. (2020)	Transparency, Explainability	–	Trust	–	–	–
Werder et al. (2022)	Transparency, Explainability	Accountability	–	–	–	–
Papers including principle	*44*	*28*	*6*	*7*	*20*	*4*
Percentage of papers with principles	*92%*	*58%*	*13%*	*15%*	*42%*	*8%*
Percentage of all papers	*81%*	*52%*	*11%*	*13%*	*37%*	*7%*

* Parentheses show principles before pilot study.
Only papers presenting a set of principles are included in this table.