Doctoral Dissertation

Design Principles for Algorithmic Accountability: an Elaborated Action Design Research

Graduate School of International Social Sciences
Yokohama National University

ALEKSANDRA TOMILOVA

June 2021

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abbreviations

A/IS	Autonomous / Intelligent Systems
ACM	Association for Computing Machinery
ADR	Action Design Research
AI	Artificial Intelligence
AS	Algorithmic Systems
BIE	Building and Intervention
BMC	Business Model Canvas
DP	Design Principle
DSR	Design Science Research
e-ADR	Elaborated Action Design Research
EAD	Ethically Aligned Design
EC	Empirical Claims
FAT	Fairness, Accountability and Transparency
ICT	Information and Communication Technology
IS	Information Systems
IT	Information Technology
MIS	Management Information Systems
MNC	Multinational Corporation
MVP	Minimum Viable Product
QDA	Qualitative Data Analysis

Abstract

Rapidly expanding application of algorithms in the workplace and our everyday lives has led to emerging new challenges related to their scrutiny and accountability. Today organizations face legal, ethical and brand reputation consequences caused by algorithmic bias and other impacts of algorithmic systems usage. This study seeks to contribute to IS literature by proposing a set of design principles for improving algorithmic accountability as a part of an organizational IT strategy. Drawing on accountability and ethically aligned design theories, this study utilizes action design research methodology based on the data gathered within the context of an immersive practice-based project. We applied an e-ADR method as our research method of choice due to identified fitness of ADR for investigation and development of socio-technical artefacts. We collaborated with practitioners and involved a number of stakeholders throughout the project, which unfolded in the context of a Japanese branch of the globally operating technology company. We constructed and evaluated Algorithmic Accountability Canvas as an artefact aimed to solve the problem of improving algorithmic accountability in an organizational context, produced learning and reflections and obtained design knowledge formalized in a set of associated design principles.

Chapter 1. Introduction

1.1 Research background

Algorithms are increasingly applied within a wide range of fields and silently influence our lives on the daily. From programmatic advertising and dynamic pricing to fraud detection, disease diagnosis to high frequency trading - these are just few of the examples of algorithms that over the years have become an integral part of our society (Kitchin, 2017; Pasquale, 2015). While the benefits of algorithms are numerous: tailored news and recommendations, more accurate predictions, lower error rates and increased efficiency, the issue of tracking and assessing how those algorithms work has become an area of public concern.

We have already witnessed the cases of algorithms "gone wrong", resulting in harmful aftereffects for organizations and people involved. Recent prominent example includes an algorithm significantly marking down the grades for 2020 "A level" exam takers in the United Kingdom, disproportionally affecting students from poorer backgrounds (Ofqual, 2020). Another widely known case is COMPAS algorithm used in the US court systems, which predicted that Black defendants were far more likely to be incorrectly judged to be at higher level of risk for recidivism, while white defendants were more likely to be incorrectly marked as low risk (Larson et al., 2016). Rising concern about algorithmic accountability and fairness can be linked to recent cases of algorithmic bias and discrimination, such as Amazon's AI-based recruiting tool that favored men applicants over women (Dastin, 2018) and predictive healthcare algorithm that was found to be biased against Black patients in the US-based hospitals (Ledford, 2019).

The role and the scope of algorithms and their utilization in our society is rapidly changing, together with the growing risks of algorithmic bias, discrimination and false information spreading. Organizations that design and deploy algorithmic systems that may have an impact on people's lives are under attention following the growing concern over accountability and transparency of such systems. While some initiatives from the government and policy makers (AI Now Institute, 2018; Donovan et al., 2018) and academia (Binns, 2018; Buhmann et al., 2020; Lee et al., 2019) already exist, trusted empirical research is still scarce. Particularly, while the issues of transparency and users' right for an explanation on how the algorithms work were attempted to be theorized (Ananny & Crawford, 2018), tangible guidelines for businesses concerning the design for accountable algorithmic systems as a part of an organizational IT strategy have not been adequately addressed as a problem domain in IS research.

1.2 Research motivation

In the pre-algorithmic era (even though algorithm does not necessarily refer to computer algorithm, in this study the specific context of algorithms as computer models making inferences from data is considered unless stated otherwise) decision making in hiring, lending, health diagnosis and many other fields were made by humans. Today the growing number of these decisions is either made or to some extent influenced by algorithmic systems, tackling huge volumes of data every second and affecting decisions of various people in a range of different tasks. The role of algorithmic systems in our society is rapidly changing, together with growing risks over spreading false information, algorithmic bias and sustaining discriminating patterns (Martin, 2019). As a result, organizations that utilize algorithmic decision-making systems that may have a significant impact on a people's lives are under attention following the growing concern over accountability of these systems.

Researchers have previously proposed several main ways to deal with algorithmic accountability issues. One of them includes making sure that certain values are set during the algorithms' development stage - that is, the tech companies themselves ensure that fairness and non-discriminating practices are implemented in the algorithmic process (Courtland, 2018) Another method includes a slightly different approach, through «feeding» the algorithm some data and observing the outcomes (Diakopoulos, 2015; Eslami et al., 2017). By analyzing how the algorithm operates and what the results are based on a variety of circumstances, further actions are taken depending on whether some kind of systematic bias had been discovered. This process is also known as an algorithmic auditing, a growing area of interest for both the academia and business practitioners (Raji et al., 2020).

Currently there is no set of solidified rules for ensuring algorithmic fairness and accountability, with legislation of different regions taking varying approaches to it. As an example, the US has taken a more «sectoral» approach, meaning that different industries have their own tactics and practices related to the issue (Clarke, 2019), while the EU legislation can be considered as more generalist.

The policy risks placed on the business models stemming from the limits on algorithmic usage is a growing area of concern. While the companies previously had to adjust for the newly emerged privacy expectations by investing into understanding the norms and ensuring compliance, the similar process currently is needed for machine learning algorithms as well.

One of the issues with algorithmic accountability initiatives and the current call for companies to be held accountable for conducting impact assessment on the algorithmic systems deployed is the fact that in a business environment with multiple groups of stakeholders involved it is difficult to achieve consistency. Accountability can be considered as a non-functional requirement in system architecture (Blum, 1992), making it difficult to measure and enforce across all the levels in an organization as well as making accountability requirements formalized and shareable among all the stakeholder groups. Therefore, a challenging task for both the practitioners and academia today is to achieve consistency by ensuring that algorithmic accountability is realized at all the organizational levels and can be captured effectively at various stages during system development.

Another important factor to be considered is the non-linear nature of software and process development. A typical software development process can involve multiple beta testing stages, as well as minor updates to existing software, adding new functionality and so on. It would not be feasible to require a company to conduct a new impact assessment for every minor software update, but the current legislation initiatives, such as the US Algorithmic Accountability Act (Clarke, 2019) does not provide recommendations on how to effectively integrate impact assessments with the software development process. This poses yet another serious challenge for tech firms deploying algorithmic systems. Therefore, a more feasible and practical framework for the businesses should be developed, taking into account the variety of procedural and technological mechanisms.

1.3 Problem statement

A growing number of organizations deploy algorithmic systems, making decisions either on behalf or assisting to some extent the human actor in the decision-making process. The resulting accountability relationship is different from accountability emerging in the human only setting due to opacity of algorithmic systems. As firms face legal, ethical and brand reputation consequences emerging from failure to design for accountable algorithmic systems, they are challenged by the growing need to introduce and sustain algorithmic accountability. However, tangible and empirically tested guidelines for organizations on how to implement algorithmic accountability, making it consistent and shareable across all stakeholders involved does not exist in practice. This study will address the problem of improving algorithmic accountability through developing a set of design principles in an organizational context.

In connection with the research problem outlined above, researcher aims to answer the following question:

- What are the appropriate design principles for improving algorithmic accountability in the organizational context?

Additionally, four sub-questions were developed in order to provide further support in addressing the main research question. In general, sub-research questions go in parallel with the four stages of Action Design Research methodology applied in this study.

- How is algorithmic accountability realized in a case organization and what factors can serve as either facilitators or barriers for achieving it?
- What are the critical design principles and features for facilitating algorithmic accountability in a case company?
- How does the instantiated artefact (set of design principles) help to solve the identified problem?
- How can a problem solution generalization for improving algorithmic accountability in an organizational context be developed?

1.4 Research scope

Rapidly expanding usage of algorithms in the workplace and our everyday lives over the last decade has attracted a lot of public attention due to increasing concerns regarding how those algorithmic systems are utilized in practice. More companies are under attention and scrutiny following changes in legislation and cases of algorithmic bias and discrimination presented in the media. As algorithmic decision-making has become widespread in a number of public systems, ranging from finance to healthcare, policing and mobility, usage of algorithmic systems in private companies has become an area of public concern as well.

At the same time, the area of algorithmic decision-making and issues relating to accountability, algorithmic opacity, data usage and ethics have sparked an active interest in the academia in the last few years (Ananny & Crawford, 2018; Binns, 2018; Brown et al., 2019; Buhmann et al., 2020; Diakopoulos, 2015; Donovan et al., 2018; Shin & Park, 2019; Warren et al., 2019). The body of academic literature continues to grow rapidly, with researchers from various fields, including but not limited to computer science, law, business management, sociology and others contributing and exploring the nascent theory relating to algorithmic accountability issues.

Despite the recognition of importance of an algorithm as a technical construct in the academia, its socio-technical dimension appears to be neglected. Wieringa (2020) calls for increased attention in future academic research to algorithmic accountability as a phenomenon of a socio-technical nature, carrying both the technical constructs as well as social and cultural aspects to it. A review of the relevant literature revealed the significance of socio-technical systems as an area of academic and practical interest (Baxter & Sommerville, 2011; Carayon, 2006; Clegg, 2000; Fox, 1995; Herrmann et al., 2007; Ropohl, 1999). Researcher aims to explore the socio-technical view of algorithmic accountability as a concept, highlighting the socially constructed aspects of this newly emerged phenomenon. Moreover, in the scope of the current study, researcher attempts to understand the nature of emerging algorithmic accountability relationships, including its distribution between different actors and across various levels within the organization.

Researcher wishes to reflect the current state of algorithmic accountability through an extensive case study by conducting an Action Design Research within the organizational context of a large MNC located in Japan. The study will aim to contribute both to theoretical and practical parts of knowledge by developing a set of design principles for achieving accountability in algorithmic decision-making processes in a business setting. The scope of the study is further defined by conditions relating to empirical setting and data collection activities, which will be realized through Action Design Research team formation. The ADR team will consist of the researcher herself and case company stakeholders, including engineering, R&D and business side associates.

1.5 Expected contribution

The study aims to contribute to both theoretical and practical streams of knowledge generation. The summary for expected contribution is presented below.

1.5.1 Methodology

The study attempts to solve an organizational problem of improving algorithmic accountability within the organizational context by building an innovative artefact (a set of design principles assisting the case organization in implementing accountable algorithmic systems) in a specific context (Japanese branch of a large multinational corporation), addressing a particular class of problems (algorithmic accountability).

Through applying ADR as a research method for this longitudinal study, researcher expects to contribute to the body of knowledge in the IS field by employing the elaborated ADR process model proposed by Mullarkey and Hevner (2019) in order to iterate nascent design theory to inform IT artefact design and use across problem domain in question (algorithmic accountability). Therefore, researcher attempts not only to inform research and practice by developing an innovative IT artefact for specific contextual use but demonstrate its utility across the whole class of field problems domain.

Moreover, researcher expects to contribute to IS theory by grounding accountability theory on the algorithmic studies and reflecting the socio-technical nature of the algorithmic accountability phenomenon.

1.5.2 Design

As this study utilizes ADR as a research method, it aims to build prescriptive design knowledge through developing and evaluating an ensemble IT artifact in the context of an organization. To show the design knowledge unfolding from the application of ADR, the conceptual framework in a form of the set of design principles for accountable algorithmic systems will be developed as a result of iterative ADR cycles and their evaluation. One of the relevant issues for researchers engaging in ADR is general solution concept formulation (generalization) due to situational nature of IT artifact development process. Therefore, this study aims to contribute to problem solution generalization and ensemble-specific knowledge creation by articulating class of problems and class of solutions.

1.5.3 Practical contribution

This longitudinal study unfolds in a real business setting within the context of a case company, a Japanese branch of a big German automotive and technology MNC. Through forming the ADR team consisting of researcher and practitioners, researcher expects to actively involve a number of relevant stakeholders, including technologists (developers, engineers), managers and other employees. This study aims to contribute to practice-based knowledge through ADR method by addressing the real business problem and producing a set of design principles for the management in order to assist case organization in designing accountable algorithmic systems and improving currently realized algorithmic accountability practices.

Moreover, researcher aims to provide a generalizable solution for designing accountable algorithmic systems for businesses deploying such systems, suitable for use outside of situational context outlined in the current study.

1.6 Dissertation outline

This dissertation is written as a monograph in six chapters. The chapters are presented in an order that is optimized for easier understanding of the study and information comprehension. Therefore, it is advised to start reading from the introduction part and then follow the chapters in their sequential order as presented in the dissertation contents. The structure of the dissertation is as follows:

The first chapter introduces research background, provides the necessary context on algorithmic role in our society and describes research motivation. It also establishes research problem and associated research questions, as well as gives an outline of expected theoretical and practical contributions.

The second chapter provides a review of the relevant academic literature related to algorithmic accountability and is structured as follows: first of all, an overview of accountability concept and the related conceptual issues is provided. Next, we review the body of literature on accountability at an organizational level and accountability as an individual-level constructs. Lastly, a definition for an algorithm, Ethically Aligned Design theory and an outline for algorithmic accountability and the way it relates to accountability theory are reviewed.

The third chapter introduces the reader to research design of the study, specifically we present the discussion on Design Science methods, Action Design Research and justification of its application in the study, empirical setting and data collection.

In the fourth chapter we present empirical part of the study by discussing the four stages of the e-ADR project, namely Diagnosis, Design, Implementation and Evolution.

The fifth chapter synthesizes learnings from e-ADR project by presenting discussion of theoretical contributions and practical implications, acknowledges limitations of the study and provides possibilities for future research.

The sixth chapter concludes the study by providing closing remarks and summing up research implications and outcomes through re-visiting the original research questions.

Chapter 2. Review of literature

Academic scholarship on accountability has a long history, with studies addressing both accountability at the organizational level (Bovens, 2007; Corts, 2007; Eisenhardt, 1989; Frink et al., 2008; Schedler, 1999) and as an individual-level construct (Dose & Klimoski, 1995; Kroon et al., 1991). Importance of topics such as accountability and moral responsibility as one of the basic principles of maintaining social systems and organizations have been recognized by Greek philosophers such as Aristotle and Plato (Roberts, 1989). Since the time of ancient philosophers, people were concerned about keeping the power under control, preventing its abuse and subjecting it to the rules of conduct (Schedler, 1999). In modern times it is the term *accountability* that tends to express concern about exercise or power and system of oversight (Schedler, 1999).

This following review of literature attempts to capture the main body of academic literature related to algorithmic accountability and is structured as follows: first of all, an overview of accountability concept and the related conceptual issues is provided. Next, the major theories and conceptualizations in regard to accountability in organizations and accountability as an individual-level construct each are given. Lastly, a definition for an algorithm, Ethically Aligned Design theory and an outline for algorithmic accountability and the way it relates to accountability theory are reviewed.

2.1 Accountability as a modern buzzword: conceptual issues

Accountability has been named the buzzword of the modern governance (Bovens et al., 2014). Bovens et al. (2014) mention that historically the word "accountability" was closely related to the concept of accounting however has since moved from its bookkeeping roots and became one of the symbols for fair governance. The nature of relationships has also changed, where in modern times it is the authorities themselves who are held accountable for their actions, rather than the original meaning of sovereigns holding the subjects accountable (Bovens et al., 2014).

The concept of accountability steadily gained recognition in the modern public and, especially, political discourse in the recent decades. Accountability became a buzzword and synonymous with many other loosely defined concepts and words, such as equity, transparency and even democracy (Bovens et al., 2014). Dubnick (2002) argues that accountability is dependent on a number of contextual and cultural factors and usually "holds the promise of bringing someone to justice, of generating desired performance through control and oversight, of promoting democracy through institutional forms, and of facilitating ethical behavior" (p.2). According to evidence from the US legislation study, the term "accountability" was applied from fifty to seventy different bills for each two-year term, and a more detailed examination revealed that the usage of the word had an extremely broad range from distinct individuals to different industries or agencies and rarely appeared more than one time within the specific bill, let alone defined (Dubnick, 2002). Three main problems related to the word accountability are defined: etymology not containing the conceptual history (even though the word "accountability" was not used until relatively recently, the concept of accountability itself has been around for many centuries), general ambiguity of accountability when it is treated as a word and not a concept and, finally, the lack of common language for efficient translation of the word "accountability" across different cultures and contexts. Dubnick (2002) provides a conceptual definition for accountability as a "form of governance that depends on the dynamic social interactions and mechanisms created within of such a moral community" (p.7).

Ambiguity of the word "accountability" was also emphasized in the psychology study of accountability impact on a range of social choices and judgments by Lerner and Tetlock (1999). Accountability here is also addressed as a "modern buzzword", as debates regarding who should answer to whom, when and under what conditions reign over the recent not only political, but also civil and criminal justice, healthcare, educational and other agendas (Lerner & Tetlock, 1999). The study suggests that it is a mistake to view accountability as a unitary phenomenon, as "even the simplest accountability manipulation necessarily implicates several empirically distinguishable submanipulations". It is important to realize the complex nature of accountability and accountability relationships, as a wide range of distinct accountability types exist between individuals and organizations (p. 255).

Unsurprisingly, the concept of accountability has been a center of attention in various studies across different academic fields, among them the most representative being social psychology, political science, public administration and law. However, most researchers use quite similar and overall comparable notions regarding what forms the core for accountability (Bovens et al., 2014). According to Bovens (2014), almost forty percent of the recent articles related to the topic of accountability use the formal definition for accountability that is fairly similar to the "minimal conceptual consensus" developed by Schillemans (2013). The minimal conceptual consensus contains four major components. First of all, accountability is about answerability and providing answers; "towards others with a legitimate claim in some agents' work" (Schillemans, 2013). Moreover, accountability also serves as a relational concept, emphasizing on agents who perform tasks and therefore are held accountable for their actions by others. The third observation from the minimal conceptual consensus states that accountability is retrospective and views behavior of the agent in general, which may range from results and performance to some standards or normatives. Finally, accountability refers not to a singular situation or moment, but rather implies a layered, complex process. In connection to the last observation, Schillemans (2013) proposes three phases for accountability, notably information phase (in which the agent provides an account of his conduct to the other party), the debating phase (where the forum assesses the given information and the parties involved then proceed to discuss the results) and, finally, the judgement phase (in which the decision about sanctions is made).

Schedler (1999) attempted to reconstruct the modern concept of accountability the way we currently use it. Accountability is presented here as a two-dimensional concept, carrying two connotations: answerability, the obligation of the officials to inform about and explain their actions; and enforcement, the capacity of accounting actors to impose sanctions. (Schedler, 1999).

When it comes to political accountability, both of the connotations are usually present, but cases when only one of the two aspects dominates (either accountability as answerability or accountability as enforcement) also exist. Schedler (1999) argues that when defining the term "accountability", people often tend to use the word "answerability" as the closest in meaning and provides a definition of accountability as "A is accountable to B when A is obliged to inform B about A's (past or future) actions and decisions, to justify them, and to suffer punishment in the case of eventual misconduct" (p.17). The concept of accountability is linked to transparency, as demand for accountability originates from the opacity of power. Accountability also notably differs from the concept of supervision, as accountability contains implications for public disclosure, whereas in case of supervision, the supervising actor can often remain invisible and unknown to the eye of the public.

In a widely recognized conceptualization for accountability proposed by Bovens (2007), accountability is seen in a narrow sense as a "relationship between an actor and the forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can pose questions and pass judgement, and the actor may face consequences» (p. 447). The study suggests that accountability can be viewed either in a broad or narrow sense, where the former refers to evaluative concept close in meaning to the word «responsiveness» and where there is no consensus on what can be considered an accountable behavior due to differences from individual to individual, time to time or place to place. Accountability in a narrow sense, however, contains two major elements to it, namely an actor (which can be either an organization, an individual or other agency) and an accountability forum (specific individual or an organization such as the parliament, court or audit office). The relationship between the two may have the principal - agent nature (Eisenhardt, 1989), where an accountability forum serves as a principal and an actor serves as an agent (Bovens, 2007). The nature of the actor - forum relationship also suggests the existence of obligation that is imposed on the actor, which can be either formal or informal (such as voluntary audits) (Bovens, 2007). The key aspect of the forum - actor relation is an obligation of an actor to inform forum regarding his or her conduct, providing the necessary data about performance or outcomes. The forum subsequently questions an actor in relation to the provided data and passes judgement regarding his or her conduct. The study suggests that the notion of sanctions should be included in conceptualization of accountability in its narrow sense, as "*possibility* of sanctions - not the actual imposition of sanctions - makes the difference between non-committal provision of information and being held to account" (p.451). Lastly, an actor may face consequences, either highly formalized or implicit.

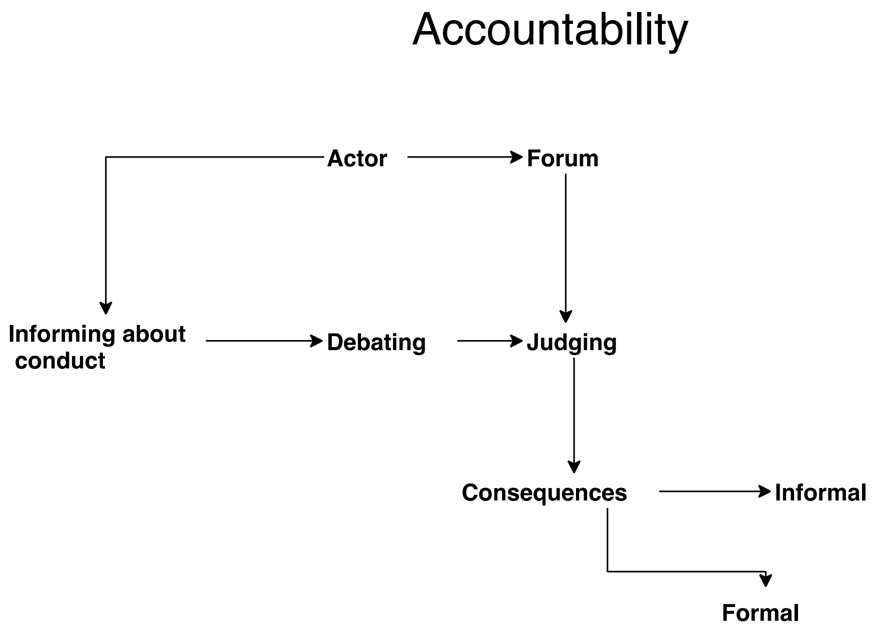


Figure 1. Conceptualization of accountability taken from Bovens (2007)

In connection to the proposed accountability conceptualization, Bovens (2007) suggests that there are four important questions in total that should be answered. First of all, it is necessary to find out to whom is account to be rendered, which will result in classifying accountability based on the type of the forum. The second question is who should render account, which leads to classifying the actor itself. The third question is about what account should be rendered, which subsequently results in classification based on the types of accountabilities. Finally, the fourth question is why the actor feels forced to render account, which yields in classification of accountability based on the nature of obligation.

Table 1. Types of accountabilities taken from Bovens (2007)

Based on the nature of the forum	*Based on the nature of the actor*	*Based on the nature of the conduct*	*Based on the nature of the obligation*
▪ Political accountability ▪ Legal accountability ▪ Administrative accountability ▪ Professional accountability	▪ Corporate accountability ▪ Hierarchical accountability ▪ Collective accountability ▪ Individual accountability	▪ Financial accountability ▪ Procedural accountability ▪ Product accountability	▪ Vertical accountability ▪ Diagonal accountability ▪ Horizontal accountability

▪ Social accountability	

To conclude, even though the term accountability has a long history and notably gained its momentum in the academia, the ambiguity of the concept and relatively sparse theoretical foundations (especially in the area of accountability as an individual-level construct) create some challenges and call for further investigation for its improved conceptualization.

2.2 Accountability in organizations

Accountability has been primarily addressed in academic research at the firm-level as opposed to its view as an individual level construct. Many studies on accountability in organizations to date are based on the agency theory (Eisenhardt, 1989), which views principal-agent relationships in organizations in the light of self-interest and incentives. In the context of accountability, the role of an agent is realized by some entity (either group, organization or an individual), the activities of which are evaluated by another party; whereas the principal refers to a role of a person or persons to observe and evaluate the aforementioned agent. The dominant perspective on accountability as a political, social and administrative mechanism suggests that accountability can be viewed as an institutional arrangement, in which an agent can be held accountable by another institution or agent (Bovens et al., 2014). It is important to note that accountability at the firm level and individual accountability research both explore the actions and behaviors of specific individuals, however in case of accountability at the firm-level, the point of interest is an organizational-level outcome, such as financial performance. Moreover, the recent academic studies on accountability also focus on individual-level accountability within the organizational context.

The issue of teams versus individual accountability and solving multitask problems through job design has been addressed in an organizational economics study conducted by Corts (2007), who has developed a multitask agency model. The study provides an argument for aligning incentive compensation in accordance with tasks of each worker and suggests that organizations based on the joint contributions of large teams "may arise endogenously as optimal organizational forms, even when there exist performance measures that reflect each worker's contribution alone" (Corts, 2007). The study emphasizes the idea that joint accountability of teams helps to mitigate multitask problems, even though it may seem paradoxical and counterintuitive, since assigning tasks to the workers this way makes the performance measures less informative for each specific agent involved.

Frink and Klimoski (1998) attempted to theorize the concept of accountability by advancing the framework of the development of shared role expectations and organizational roles (Katz & Kahn, 1978). Frink and Klimoski (1998) argue that accountability theory in principle is based on similar explanations for predictable behavior as the role theory, which describes how organizations manage to produce reliable behaviors on the part of their members. Both accountability theory and the role theory emphasize the importance of interpersonal relationships. Moreover, the study suggests that both role and accountability theories stress the importance of interpersonal expectations and connect activities and tasks to specific individuals (Cummings & Anton, 1990; Schlenker et al., 1994). To summarize, accountability theory can be viewed as containing aspects of role making and role taking in the context of role episodes as also suggested by the role theory (Frink & Klimoski, 2004). Frank and Klimoski (1998, 2004) also argue that adopting a role theory perspective for theorizing accountability has several advantages, such as a different unit of analysis (here "relationship" suggested as the optimal unit of analysis instead of "event"), multiple set of expectations, the dark side of accountability (necessity to admit some potential undesirable effects that can occur either organizationally or socially), and the dynamics of accountability contexts (incorporating a variety of intrapersonal, interpersonal and person-organization dynamics as multiple factors in a unified framework).

A study on organizational control systems by Dose and Klimoski (1995) highlights the issue of internal versus external control in accountability theory. The study proposes a progressive view for accountability theory, in which external control needed to be realized by organizations and internal control realized by employees' self-management practices and felt responsibility could effectively coexist. The proposed view of coexistence between accountability as a social control and responsibility as a self-control suggests that accountability as an element of organizations' external control system can actually increase agents' (employee) internal control by enhancing individual feelings of responsibility. A conventional view of accountability as a form of being held accountable through some variation of a reporting requirement is not efficient and will only lead to dysfunctions within the traditional external control system of the organization. Dose and Klimoski (1995) link accountability with an identity theory (Schlenker, 1986) which is based on three elements - events, prescriptions and identities. Accountability is related to a particular event - such as outcome or resulting performance; identity refers to agent's self-concept; prescription refers to behavior standards associated with a particular event. Altogether these elements provide an explanation for the impact of accountability on the specific individual (Dose & Klimoski, 1995; Schlenker & Weigold, 1992). The study suggests that the linkages between those three elements (events, prescriptions and identities) are key factors in predicting accountability effect, as the stronger the linkages are, the greater is the impact for the individual and, subsequently, the strength of the accountability force. Based on the dynamics described above, Dose and Klimoski (1995) proposed a progressive accountability model (see Figure 2), framing accountability discussion in terms of identity theory and suggesting that accountability as a social control element can have a positive effect on self-control (felt responsibility) of an agent.

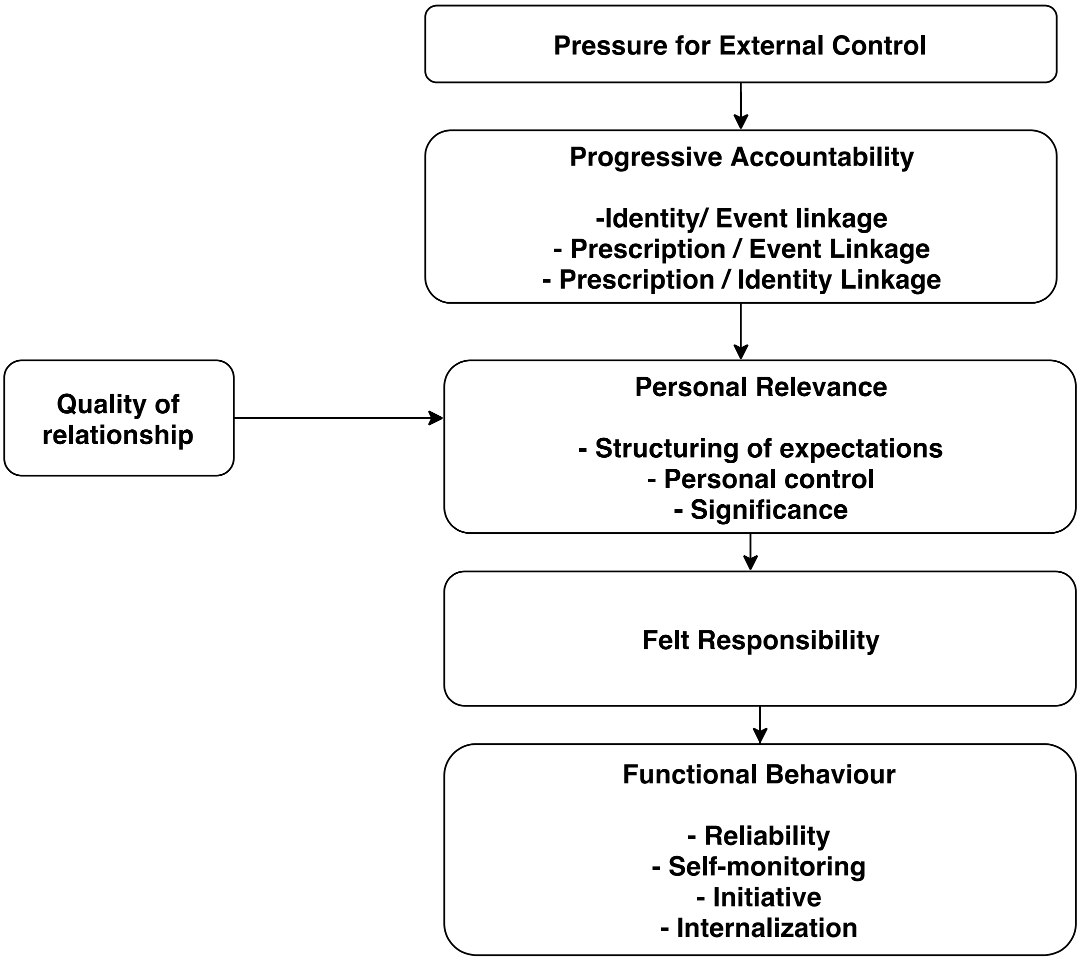


Figure 2. Progressive accountability model taken from Dose and Klimoski (1995)

Frink et al. (2008) provide a meso-level conceptualization of accountability, stressing the importance of multilevel research for accountability as one of the key elements for modern organizational success and establishing its antecedents and outcomes as an area of great importance for both the practitioners and academics. The study argues that because of the dynamic, multi-level nature of organizations, single-level conceptualizations for accountability are not sufficient and at times even misleading. Even though enacting accountability within organization involves evaluation at some stage, the critical point for accountability is not centered around evaluation itself, but answerability (Frink et al., 2008). Moreover, implementing formal requirements for answering is not a necessity for enacting accountability, but rather it is a perception that it may occur that calls on its effect. The suggested meso-level conceptualization model for accountability contains eight different elements: environmental factors, accountability systems, features of the accountability environment, the experience of accountability, resources and capabilities, reputation, performance and well-being (Frink et al., 2008). The proposed model aimed to develop a more comprehensive understanding of accountability as a concept, to integrate the recent research and theory and to expand a single, unitary perspective regarding accountability towards a more holistic meso-level conceptualization.

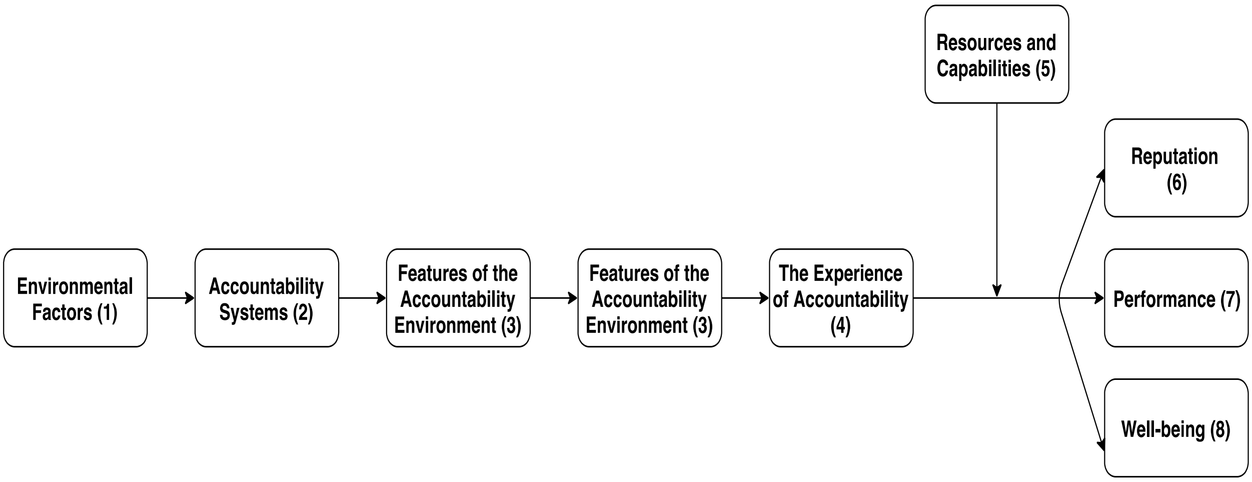

Resources and Capabilities (5)
Reputation (6)
Environmental Factors (1)
Accountability Systems (2)
Features of the Accountability Environment (3)
Features of the Accountability Environment (3)
The Experience of Accountability (4)
Performance (7)
Well-being (8)

Figure 3.Meso-level conceptualization of accountability taken from Frink et al. (2008)

To sum up, accountability in organizations has been primarily addressed at the firm-level by investigating the area of corporate governance and is mainly based on the agency theory (Eisenhardt, 1989). In the following part of the study we would like to address the extant literature on accountability as an individual-level construct.

2.3 Accountability as an individual level construct

A separate stream in academic scholarship on accountability addresses it as an individual-level construct, often referred to as felt accountability or simply accountability. Felt accountability relates to perceptions of accountability of the actor (Frink & Klimoski, 1998) in contrast with attributions of accountability that were imposed on the actor by the forum (Hall et al., 2017). Lerner and Tetlock (1999) argue that despite a widespread attention to the concept of accountability in many fields in the recent decades, historically psychological research on accountability has been quite sparce. More recently there has been a growing interest to felt accountability in social psychology research (Hall et al., 2017; Hochwarter et al., 2005; Laird et al., 2015; Mackey et al., 2018; Royle, M. Todd and Hall, 2012), addressing topics and accountability relationships such as external stressors, job tension, personal reputation, entitlement and theory of needs. However, most of the contemporary research addressing felt accountability can be linked to accountability conceptualizations based on Cummings and Anton (1990), Schlenker (1994) and Tetlock (1985, 1992).

Tetlock (1992) proposed a social contingency theory in his study on the impact of accountability on judgement and choice. The study reviewed major strategies that people use in their lives when coping with demands for accountability as well as situational and personality moderators on these strategies. The social contingency model draws on anthropological and social theory in relation to the necessary conditions for the social order in addressing accountability as a universal feature in decision-making environments (Tetlock, 1992). The social contingency theory then proposes three distinct accountability coping strategies in relation to judgement and choice (the acceptability heuristic, preemptive self-criticism and the rationalization heuristic) as well as conditions under which they are most likely to be effective and advantages and disadvantages of each one (Tetlock, 1992). Accountability serves as one of the key social contingencies that influences the behavior and actions of people, as individuals are concerned about their impression, social image and identity. The study refers to symbolic motives as a part of theories of impression management and self-esteem maintenance, among which the most important are motivation to protect and enhance one’s social image and identity, the motivation to acquire power and wealth and motivation to protect and enhance one’s self-image. Compared to the pyramid model of accountability by Schlenker (1994), discussed in the next paragraph, Tetlock (1992) fully emphasized on psychology of accountability (such as internal coping strategies of an individual and psychological processes). The study also proposed phenomenological view of accountability, which focuses on individual’s subjective interpretations of accountability rather than objective mechanisms of accountability that may be formally imposed on an individual (state of mind rather than state of affairs accountability view) (Hall et al., 2017).

The pyramid model of accountability proposed by Schlenker (1994) is one of the key models attempting to conceptualize individual accountability. The study suggests that responsibility serves as a “necessary component of the process of holding people accountable for their conduct” (p.634) and refers to accountability as a “mechanism through which societies can control conduct of their members” (p.634). The pyramid model of accountability posits that responsibility serves as a key concept in understanding how people view and control each other’s conduct (Schlenker et al., 1994). Accountability contains an evaluative reckoning on the basis of which individuals are judged. The proposed evaluative reckoning consists of three elements, namely prescriptions, the event and set of identity images. In the pyramid model of accountability (Figure 4) prescriptions refer to specific guide of conduct for the individual’s actions, the events refer to specific occasion that has either already occurred or is anticipated in relation to prescriptions and the set of identity images that describe an actor’s qualities, roles and aspirations relevant to situational context. The study supported three layer model, showing evidence that attributions of responsibility are direct function of the combined linkages (prescription, event and identity) and that when judging responsibility, individuals seek for information that is relevant to those three linkages (Schlenker et al., 1994). The proposed model also contributed to clarification of responsibility concept and provided a framework for understanding social judgement.

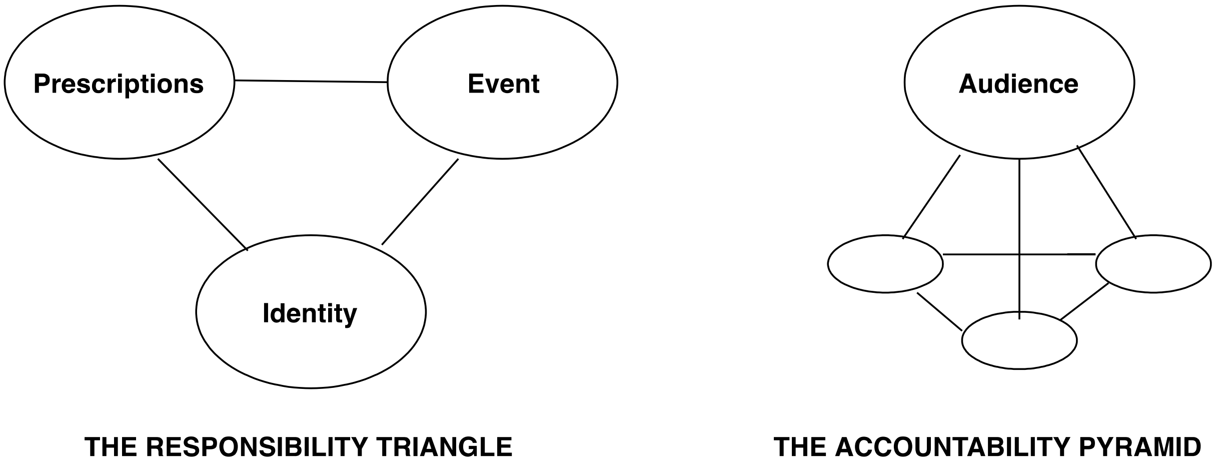


Figure 4. The responsibility triangle (left) and the accountability pyramid (right) taken from Schlenker (1994)

Hochwarter et al. (2005) describe a personality variable named Negative Affinity (NA), referring to the extent of which an individual experiences anger, anxiety, fear or hostility (D. Watson & Clark, 1984) and examine negative affinity as the moderator of the form of the relationship between felt accountability and job tension. The study contributed to accountability, stress and job tension research by demonstrating that felt accountability can either positively or negatively predict job tension and providing evidence that increased levels of perceived accountability may lead to increased tension in individuals.

A recent study by Hall et al. (2017) attempted to synthesize empirical and theoretical up to date research on felt accountability. The study aimed to provide a comprehensive review of key theories that founded contemporary body of research on felt accountability since Lerner and Tetlock (1999) and also describe an agenda for empirical studies on the same topic. The study outlined the key hurdles for felt accountability research, including identified gap, limitations and suggestions for scholars interested in making contributions to academic research on felt accountability in the future. Hall et al. (2017) argue that even though importance of accountability has been highlighted in many areas, accountability as a research domain still remains in the nascent stage and many aspects about accountability as a construct remain uncovered.

2.4 The concept of an algorithm

Many aspects in our everyday lives are being influenced and regulated by various software-enabled technologies. The software is fundamentally composed of algorithms - sets of defined steps structured to process data to produce an output (Kitchin, 2017). Algorithms play an important role in selecting what information is relevant to us by making recommendations and decisions, highlighting or excluding the news, managing our interactions and governing the major flows of information. As people around the world come into contact with various algorithms on a daily basis, we can observe a resurgence of interest in algorithmic studies not only from a strictly technical view, but also from social, economic, philosophical, contextual, ethical and other standpoints.

The traditional perspective of an algorithm is that of a technical construct, however algorithms do not need to be software specifically. Gillespie (2014) argues that in a broad sense an algorithm can be regarded as "encoded procedures for transforming input data into a desired output, based on specified calculations" (p.1), where the procedures determine both the problem and the steps that should be taken in for the problem to be solved. Algorithms therefore can be carried out not only by machines, but also by nature and people - an example of that would be a person following the recipe to cook a dinner or pupils learning long division in grade school (Diakopoulos, 2014). Contemporary academic research in the field of algorithms is focused on algorithms carried out by machines, where in a colloquial sense algorithm refers to some kind of instructions fed to a computer. However, algorithmic systems are a technical construct that carries social and cultural aspects to it (Wieringa, 2020), therefore it is important to approach algorithmic systems as a kind of system that consists of both technical and social elements. The system that is comprised of both social and technical elements can be regarded as "sociotechnical system" (Selbst et al., 2019). A rapidly growing body of research on sociotechnical systems in the recent years clearly demonstrates increasing relevance of sociotechnical systems as a phenomenon (Baxter & Sommerville, 2011; Carayon, 2006; Clegg, 2000; Fox, 1995; Herrmann et al., 2007; Ropohl, 1999). Algorithms that are used today are not just programming code with some kind of consequence, but also the new socially constructed and institutionally managed mechanism and a new knowledge logic (Gillespie, 2014).

There were many attempts to formalize the concept of algorithm in academic research throughout the years. Computer science, for which algorithm is one of the major concepts, has raised many key questions regarding algorithm as a technical construct. Historically, the concept of algorithm occupies the central place in computer science due to the way it translates the basic logic behind the Turing machine (Matthew, 2008). The word "algorithm" itself has been derived from a mixture of words "arithmos" (meaning a "number" in Greek) and "algorism", which used to refer to the art of calculating using Arabic numerals during the Middle Ages (Marciszewski, 1981). Specifically, the word "algorism" is derived from the name of Arabian mathematician

Muḥammad ibn Mūsā al-Khwārizmī (Miyazaki, 2012). Kowalski (1979) argues that conventional algorithm can be regarded as consisting of two components: a logical component, which specifies what needs to be done and the knowledge to be used in solving problems; and control component, which specifies how it is to be done and problem-solving strategies in using the knowledge. Changing the control aspect of an algorithm without altering the logic of it can therefore increase the efficiency of the algorithm (Kowalski, 1979).

Kitchin (2017) provided an overview of modern algorithmic studies and the notion of algorithm, attempting to synthesize and extend critical thinking about algorithms and the ways to research them in practice. Lack of research on algorithmic systems from a critical humanities and social sciences perspective is pointed out as opposed to vast body of literature on algorithms from a purely technological view. Moreover, three major obstacles for the emerging research in algorithmic systems were identified: gaining an access to the relevant information and algorithms' formulation (black box nature of algorithms), algorithms' heterogeneity and the way they unfold contextually and contingently. Finally, the study developed six distinct methodological approaches for researching algorithms. Examining pseudo-source and/or source code by deconstruction, examining documentation, mapping out a genealogy of how algorithm evolves and changes over time and inspecting how the same task is translated to different software languages and acts on different platforms is suggested as the first approach (Kitchin, 2017). The second methodological approach for researching algorithms deals with reflexively producing the code, in which researcher interrogates their own experience of formulating an algorithm rather than examining an algorithm produced by others (Kitchin, 2017). The third approach suggests reverse engineering, defined as "the process of articulating the specifications of a system through a rigorous examination drawing on domain knowledge, observation, and deduction to unearth a model of how that system works" (Kitchin, 2017, p. 23). The fourth approach proposed is interviewing designer of conducting an ethnography of a coding team to reveal the story behind the algorithm's production and the fifth approach for researching algorithms is unpacking the full socio-technical perspective of their creation, including wider institutional view such as legal frameworks, management, institutions and other elements. Lastly, the sixth approach is about examining how algorithms work in the world in general by inspecting how they are used and perform a variety of tasks within different domains.

2.5 Ethically Aligned Design

The issue of ethics in relation to AI has sparked an active interest in ICT research (Buhmann et al., 2020; Martin, 2019). As algorithmic decision-making becomes widespread in a number of publicly accessible systems, ranging from finance to healthcare, policing and mobility, more attention is being paid to how those algorithms operate (Pasquale, 2015). Moreover, utilization of algorithms involves numerous ethical considerations at the design, development and deployment stages of their lifecycle (Binns, 2018).

The newly emerging social and ethical implications of using algorithms have recently been addressed as Ethically Aligned Design (EAD), which refers to alignment of autonomous and intelligent technical systems (A/IS) design with values and ethical principles of the society (EAD1e, 2019). EAD addresses a broad range of issues related to human-centric design and A/IS for sustainable development, embedding values into A/IS and ethical due diligence for corporations and technologists (Weng & Hirata, 2018). The principles of EAD consider both the role of the A/IS creator, operators and any other affected parties or stakeholders (EAD1e, 2019). Accountability serves as one of the fundamental principles for ethically aligned A/IS and lack of accountability is acknowledged to present a major challenge for the implementation of the A/IS development and application (EAD1e, 2019; Vakkuri & Abrahamsson, 2018).

The field of EAD aims to reflect anthropological, political and technical aspects of applying A/IS and align their design with the values and needs of the society (Vakkuri & Abrahamsson, 2018). As ethics and sustainable development of A/IS becomes a focal point of interest for both the practitioners and academia, the discussion within EAD field is expected to grow in the near future.

2.6 Connecting algorithms with accountability theory: the concept of algorithmic accountability

As algorithms are increasingly applied across various fields and industries, affecting the lives of various people every day, it becomes crucial to track and assess how algorithms work, including identifying prejudices and bias potentially resulting from their application. Essentially, algorithms can no longer be considered a niche subject for programmers and scientists, but rather became a major issue of public interest. As decision-makers have to provide justification behind the results produced by algorithms used in their respective organizations, many questions regarding accountability issues remain unanswered. Companies that utilize algorithmic decision-making systems are under attention following the growing concern over possibility of algorithmic bias and discrimination, transparency of data and resulting potential reputational and other damages for the company itself.

Although originally "algorithmic accountability" as a term was coined by Diakopoulos (2013), the underlying principle behind it is not new and can be traced as far back as the emergence of automated systems. The rapidly growing body of academic studies in algorithmic accountability demonstrates active interest of researchers to this newly emerged and highly relevant area (Ananny & Crawford, 2018; Binns, 2018; Brown et al., 2019; Buhmann et al., 2020; Diakopoulos, 2015; Donovan et al., 2018; Shin & Park, 2019; Warren et al., 2019).

Ensuring the quality of algorithmic decision-making becomes an area of growing public concern following the dramatically expanding usage of algorithms in the workplace and our everyday lives. The recent surge in big data, as well as in complexity of algorithms applied in organizations has led to difficulties in securing quality assurance related to algorithmic systems' usage (Kemper & Kolkman, 2019). Nowadays algorithmic decision-making is embedded in a number of public systems, ranging from healthcare to finance, transport and policing, which calls for an increased attention and demands towards algorithmic transparency (Ananny & Crawford, 2018). In connection to this many researchers are focusing on transparency as one of the requisites for algorithmic accountability (Garfinkel et al., 2017; Kemper & Kolkman, 2019; Kizilcec, 2016; Lepri et al., 2018; Rader et al., 2018). However, recently transparency has become a subject of criticism and debates following some of the identified limitations. A study by Ananny and Crawford (2018) suggests that transparency alone cannot create accountable algorithmic systems and lists ten limitations of transparency ideal, among which are ambiguous connection to building trust, entailing professional boundary work, technical limitations, privilege of seeing over understanding, temporal limitations and others.

A recent study by Wieringa (2020) provided the most comprehensive systematic literature review on algorithmic accountability up to date, assessing 242 articles related to the topic and closely related areas such as regulation of algorithms, ethics and AI and others. Moreover, the study links algorithmic accountability to actor-forum accountability conceptualization proposed by Bovens (2007), which is discussed earlier in this study. Wieringa (2020) points out the vague nature of algorithmic accountability as a term and proposes the following definition:

> Algorithmic accountability concerns a networked account for a socio-technical algorithmic system, following the various stages of the system's lifecycle. In this accountability relationship, multiple actors (e.g., decision makers, developers, users) have the obligation to explain and justify their use, design, and/or decisions of/concerning the system and the subsequent effects of that conduct. As different kinds of actors are in play during the life of the system, they may be held to account by various types of fora (e.g., internal/external to the organization, formal/informal), either for particular aspects of the system (i.e. a modular account) or for the entirety of the system (i.e. an integral account). Such fora must be able to pose questions and pass judgement, after which one or several actors may face consequences. The relationship(s) between forum/fora and actor(s) departs from a particular perspective on accountability. (p. 10)

As one can observe from the proposed algorithmic accountability definition, it closely follows the structure and logic behind actor-forum relationship in accountability formalization described by Bovens (2007). According to Wieringa (2020), algorithmic accountability is distributed between different actors, and it is important to specify their levels, roles and type of responsibility to understand the nature of accountability relationship. Subsequently, algorithmic accountability contains different fora (as opposed to one single forum) and it is crucial to determine what each forum needs. Moreover, it includes the account at various stages in an algorithm's lifecycle in a form of ex ante, in medias res and ex post considerations (Wieringa, 2020). Similarly to accountability conceptualization in Bovens (2007), the study suggests that algorithmic accountability also includes consequences that can be imposed on an actor by the forum. Finally, the fifth element in the algorithmic accountability definition proposed is a consideration of perspective on accountability arrangement to determine what needs to be accounted for in an algorithmic accountability relationship. Wieringa (2020) calls for further investigation and integration of algorithmic studies into accountability theory, as the current research sparsely addresses the aforementioned fields. Lastly, it is argued that further studies in algorithmic accountability should address the perspective of algorithmic accountability as a sociotechnical and interdisciplinary phenomenon, as neither law, data science, governance studies nor any other field can embrace and tackle algorithmic accountability alone.

The issue of algorithmic accountability mainly in the field of media and computational journalism has been extensively covered by Diakopoulos (2015, 2016, 2017). The key component of algorithmic power is autonomous decision-making (Diakopoulos, 2014). Diakopolous (2015) suggests that algorithmic power can be assessed by investigating decisions that the algorithms make, namely classification, prioritization, association and filtering. Prioritization deals with criteria used to define some kind of ranking through a sorting procedure, classification means categorizing a particular element as a constituent of a given class, association marks relationships between entities and finally, filtering involves either including or excluding specific information due to criteria or rules (Diakopoulos, 2015). It is argued that even though algorithms exert power from the standpoint of four criteria listed above, there is still a notable range of human influences embedded in algorithms, such as training data. Therefore, it is important to refer to algorithms as the products of human development and consider their intent and the agency of actors interpreting the results produced by those algorithms in order to make higher-level decisions (Diakopoulos, 2015). In the news and media industry, many companies are increasingly using various types of algorithms in production of content, fostering the transparency debate (Diakopoulos & Koliska, 2017). Diakopoulos & Koliska (2017) also develop the summary for transparency factors for algorithmic systems and call for their evaluation in future research.

Layer	Factors
Data	• Information quality. • Accuracy. • Uncertainty (e.g. error margins). • Timeliness. • Completeness. • Sampling method. • Definitions of variables. • Provenance (e.g. sources, public or private). • Volume of training data used in machine learning. • Assumptions of data collection. • Inclusion of personally identifiable information.
Model	• Input variables and features. • Target variable(s) for optimization. • Feature weightings. • Name or type of model. • Software modeling tools used. • Source code or pseudo-code. • Ongoing human influence and updates. • Explicitly embedded rules (e.g. thresholds).
Inference	• Existence and types of inferences made. • Benchmarks for accuracy. • Error analysis (including e.g. remediation standards). • Confidence values or other uncertainty information.
Interface	• Algorithmic presence signal.

- On/off.
- Tweakability of inputs, weights.

Table 2. Summary of transparency factors across four layers of algorithmic systems taken from Diakopoulos & Koliska, 2017

In connection with the discussion above, Diakopolous (2015) points out weakness in transparency approach and suggests reverse engineering method as the more feasible way to ensure algorithmic accountability. According to the study, transparency is "far from a complete solution to balancing algorithmic power" (p.403) and a few reasons to support the argument are addressed. Firstly, transparency can only be considered useful when there is a enough motive from the side of an algorithm creator to disclose information (for example, it proved to be efficient for transparency policies like restaurant scores and automotive safety tests due to competitive dynamics and public concern for the companies involved). However, in other cases algorithm creators or operators may have some kind of conflict with transparency goals (Diakopoulos, 2015). For example, in many cases organizations may limit how transparent they are due to their concern not to disclose too many details of their systems that either may hurt their competitive advantage or hurt their reputation and potential business opportunities in any way.

Authors	Paper title	Research objectives	Context	Methodology	Main findings
Diakopoulos (2015)	Algorithmic accountability: Journalistic investigation of computational power structures	To understand the opportunities and limitations of a reverse engineering approach to investigating algorithms	Computational journalism	Case study, conceptual	Reverse engineering the input–output relationship of an algorithm was found to elucidate significant aspects of algorithms such as censorship
Ananny and Crawford (2016)	Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability	To investigate transparency limitations and to sketch an alternative typology of algorithmic accountability	Platforms and data systems	Conceptual	Typology of transparency limitations proposed; transparency is argued to be an inadequate way to govern algorithmic systems
Buhmann et al. (2019)	Managing algorithmic accountability: Balancing reputational concerns,	To suggest a framework for managing algorithmic accountability	Private and public firms, utilizing algorithms	Conceptual	Proposed discourse-ethical approach for managing opaque

	engagement strategies and the potential of rational discourse				algorithms, framework created
Binns (2018)	Algorithmic accountability and public reason	To present an account of algorithmic accountability in terms of the democratic ideal of 'public reason'	Political philosophy	Conceptual	Public reason is argued to provide a partial answer to algorithmic accountability issues
Wieringa (2020)	What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability	To provide a systematic literature review on algorithmic accountability	Socio-technical systems	Conceptual	Specified the definition for algorithmic accountability, linked it to actor-forum accountability conceptualization proposed by Bovens (2007)

Table 3. Selected conceptual studies on algorithmic accountability (2015-2020)

Table 3 represents selected studies on algorithmic accountability. The studies have been chosen based on the following criteria: first, they are conceptual articles discussing algorithmic accountability as a phenomenon of socio-technical nature. Secondly, they are selected from reputed journals such as *Journal of Business Ethics, Philosophy&Technology* and *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT)*, the largest computer science conference bringing together researchers and practitioners interested in the issues of accountability, fairness and transparency in the socio-technical domain. The articles span the period from 2015, which approximately corresponds to the earliest mentions of algorithmic accountability as a term, which was coined by Diakopoulos (2013), up until 2020.

Authors	**Paper title**	**Context**	**Method**	**Sample**
Brown et al. (2019)	Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services.	Public service agencies, child welfare	Participatory design	Fours samples: families, frontline providers, specialists, prototype specialists (n=18, 38, 11, 16)

Veale et al. (2018)	Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making	Public administration, predictive policing	Open-ended work in policy research	27 public sector machine learning practitioners across 5 OECD countries
Katell et al. (2020)	Toward Situated Interventions for Algorithmic Equity: Lessons from the Field	Co-developing algorithmic accountability interventions	Participatory and co-design methods	Community groups, civil rights organizations and advocates
Young et al. (2019)	Municipal surveillance regulation and algorithmic accountability	Surveillance practices	In-depth case study	28 surveillance technologies disclosed by municipal departments
Neyland (2016)	Bearing Accountable Witness to the Ethical Algorithmic System	STS sensibilities, ethnomethodological work on sense-making accounts	Case study	1 large technology firm, 2 large transport firms, 1 consultancy firm

Table 4. Selected empirical studies on algorithmic accountability (2015-2020)

Table 4 represents selected empirical studies on algorithmic accountability ranging from 2015 until 2020 in publication date. Analysis of extant literature revealed a noticeable lack of empirical research related to algorithmic accountability, while most of the studies focus on conceptual issues in the problem domain, such as political philosophy, law and social studies. A number of studies have employed participatory methods in the field of public sector algorithmic decision-making and surveillance technologies. However, despite identified fitness of ADR methods for developing socio-technical design agenda for a specific class of problems (Sein et al., 2011), no prior studies have utilized ADR approach in investigating algorithmic accountability. Moreover, extant research primarily attends to algorithmic decision-making and A/IS in the public services problem domain, while private sector and the issue of providing guidance on algorithmic systems design for the businesses remain to be neglected. Lastly, despite recognition of importance of an algorithm as a technical construct in the academia, its socio-technical dimension appears to be neglected. Recent studies (e.g., Wieringa, 2020; Martin, 2019) call for increased attention in future academic research to algorithmic accountability as a phenomenon of a socio-technical nature, carrying both the technical constructs as well as social and cultural aspects to it.

To conclude, algorithmic accountability as a concept and research field is still in the nascent stage, even though historically the underlying principles behind it can be traced back to decades ago. The current issues for researchers interested in examining algorithmic accountability include but not limited to reflecting its' socio-technical nature and grounding accountability theory on the algorithmic studies.

Chapter 3. Research design

3.1 Design Science methods

Design science research has been defined as «a research paradigm in which a designer answers questions relevant to human problems via the creation of innovative artefacts, thereby contributing to new knowledge of the body of scientific evidence. The designed artefacts are both useful and fundamental in understanding that problem» (Recker, 2012). Design science research is a relatively new approach and has gained much attention since the early 2000s, while the starting point is believed to be an article published by Alan Hevner et al. in the MIS Quarterly (Hevner et al., 2004).

The main idea behind design science research is understanding and acknowledging of a 1) design problem and 2) its solution in the process of building and application of an artifact. The definition of an artifact is central to IS research in general and design science research in particular and is used to describe something that is artificial (i.e., created by humans). Hevner et al. (2004) distinguishes between two complementary, but still distinct paradigms in Information Systems discipline, namely behavioral science and design science. While behavioral science mainly deals with predicting human and organizational phenomena, design science research is a problem-solving paradigm that «seeks to create innovations that define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, management, and use of information systems can be effectively and efficiently accomplished» (Hevner et al., 2004, p.76). Moreover, it is necessary to conduct complementary research between behavioural science and design science paradigms in order to address relevant problems in productive application of information technology (Hevner, 2007; Hevner et al., 2004). Hevner et al. (2004) argue that behavior and technology are inseparable in Information Systems research and call for synergistic approach in future research between the two IS paradigms.

As design science research is concerned with IT artifacts, this view is well fitted with the object of the current study, which is algorithmic accountability and the usage of algorithmic systems within the organization. In design science IT artifacts are created and evaluated to solve identified organizational problems and «the further evaluation of a new artifact in a given organizational context affords the opportunity to apply empirical and qualitative methods» (Hevner et al., 2004, p.77). In addition, in design science artifacts are created to address unsolved problems and further evaluated based on the utility that they provide by solving those problems. In this context, the newly created artifact is a tool developed to improve algorithmic accountability in the organization and the utility it demonstrates for the firm as well as the end users, as «utility can be a performance metric that defines the extent of improvement of the novel artifact over an existing solution…but also may be interpreted by end users, or in terms of efficacy, efficiency, effectiveness or other criteria» (Recker, 2012). Table 5 represents guidelines for DSR taken from Hevner et al. (2004) and corresponding research compliance.

Guideline	Research Compliance	Description
1. Design as an artefact	Creating an artefact assisting organizations in improving algorithmic accountability	Design science research must produce a viable artefact in the form of a construct, model, a method, or an instantiation
2. Problem relevance	Research is driven by identified need to introduce a comprehensive tool for organizations taking steps in achieving accountable algorithmic decision-making processes	The objective of design science research is to develop technology-based solutions to important and relevant business problems
3. Design valuation	Design artifact will be validated and evaluated through developing an artefact prototype in a form of Algorithmic Accountability Canvas	The utility, quality, and efficacy of a design artefact must be rigorously demonstrated via well-executed evaluation methods
4. Research contributions	A situated artifact to be used in an organizational setting will be developed to solve a specific identified problem in a real practical context	Effective design science research must provide transparent and verifiable contributions in the areas of the design artefact, design foundations, and/or design methodologies
5. Research rigour	Research design is guided by literature on evaluation methods from IS design science research and, specifically, action design research (ADR, e-ADR) as a subvariant of DSR	Design science research relies upon the application of rigorous methods in both the in the required fields construction and evaluation of the design artefact
6. Design as a search process	Research activities are realized in an iterative manner, engaging a number of representative stakeholders to build and evaluate an ensemble artifact	The search for a useful artefact requires utilizing available means to reach desired ends while satisfying laws in the problem environment
7. Communication of research	Both the design project phases and research findings will be communicated to stakeholders from the case organization	Design science research must be explained effectively to both technology-oriented and management-oriented audiences

Table 5. Guidelines for DSR research copied from Hevner et al., 2004 and proposed research compliance

3.2 Action Design Research

To address the outlined research questions, researcher has applied Action Design Research (ADR) methodology. ADR has been defined as a "research method for generating prescriptive design knowledge through building and evaluating ensemble IT artifacts in an organizational setting" (Sein et al., 2011, p. 40). ADR was recognized as a suitable methodology to conduct this study due to several reasons. First of all, algorithmic accountability is a concept of a socio-technical nature, calling for an interdisciplinary approach in its investigation as opposed to single-sided view such as law or computer science (Wieringa, 2020). Moreover, ADR puts a great emphasis on collaboration with practitioners and inclusion of various stakeholders, therefore proving a great fit for the phenomenon in this study that unfolds within an organizational setting. However, in Design Science Research (DSR) organizational intervention is recognized as an area of secondary importance compared to the main techno-centric view of an innovative IT artefact building and the utility it provides. Some researchers also reflected on the notion of an IT artefact and the risks associated with the narrow, techno-centric design and argue that in cases where IT artefact is not well linked to the social context, some unforeseen results may arise (Goldkuhl, 2013; Purao, 2013; Silver and Markus, 2013). In connection to the points mentioned above, Design Science Research (DSR) alone was deemed as an insufficient method to conduct this study.

On the other hand, Action Research is widely known as a method strongly oriented towards collaboration as an iterative process involving both researchers and subjects (Myers, 2009). However, Action Research does not explicitly focus on designing and building an innovative IT artefact, which is one of the key proposed components in the current study. In the seminal article introducing Action Design Research, Sein et al. (2011) argue that relevance challenge for IS calls for a research method that would recognize IT artifacts as shaped by a variety of stakeholders (such as users, investors and developers) "without letting go of the essence of design research (DR): 1) innovation and 2) dealing with a class of problems and systems" (p.38). Based on the points discussed above, ADR was chosen as a best fitting research method to address the problem posed in this study.

The current study will be validated within the context of a large multinational enterprise. As the case organization applies APM (agile project management) for its software development processes, it is necessary to indicate the fitness of action design research to agile practices. Researchers have previously pointed out the connection between action research or action design research and agile (Keijzer-Broers & de Reuver, 2016; Senabre Hidalgo & Fuster Morell, 2019). Action research is widely known to stress the importance of iterative processes and collaboration between various involved stakeholders, while APM can be considered a co-creation practice aimed to achieve adaptive, responsible teamwork through frequent and small releases and team practices such as "standup" feedback meetings and workflow visualization (Senabre Hidalgo & Fuster Morell, 2019). Keijzer-Broers & de Reuver (2016) illustrated a succesful example of combining action design research with agile and sprint methods in their study on prototyping a wellbeing platform and argue that design sprint can set up a design process in the context of Action Design Research, advancing it from the prototype phase into an MVP.

Stages and Principles		**Artifact**
Stage 1. Problem Formulation		
Principle 1. Practice-Inspired research	Research was driven by the identified need to assist organizations in improving algorithmic accountability	**Recognition:** Shortcomings of the existing practices to ensure algorithmic accountability
Principle 2. Theory-ingrained Artifact	Kernel theories identified: algorithmic accountability theory (Diakopoulos, 2015), Ethically Aligned Design theory (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019)	
Stage 2: BIE (Building, Intervention, Evaluation)		
Principle 3. Reciprocal Sharing	Problems encountered to be iteratively addressed and formulated as early design principles in collaboration with practitioners.	**Alpha Version:** The artifact conceived as a design idea, reflective of reviewed literature and findings from the initial stage of the ADR project. Evolved from a Algorithmic Accountability Canvas prototype designed by ADR team **Beta Version:** prototype solidification
Principle 4. Mutually Influential Roles	The ADR team to include researcher and practitioners in order to embody theoretical, technical, and practical perspectives.	
Principle 5. Authentic and Concurrent Evaluation	Artifact to be evaluated within the ADR team in an organizational setting	
Stage 3. Reflection and Learning		

Principle 6. Guided Emergence	The ensemble nature of the artifact to be recognized. Furthermore, artifact revisions to be considered.	**Emerging Vision and Realization** New requirements for the artifact based on results emerging in the BIE stage. A revised version of the initial design principles.
Stage 4. Formalization of Learning		
Principle 7. Generalized Outcomes	A set of design principles to assist organizations in improving algorithmic accountability to be articulated	**Ensemble Version** An ensemble embodying design principles and managerial policies for improving algorithmic accountability in the organizational context

Table 6. Summary of the ADR process in the proposed project based on Sein et al. (2011)

We outlined the summary of an ADR process based on the work of Sein et al. (2011) in Table 6. Seven principles of an ADR process are listed along with the associated description and artifact design stages.

In the scope of the study, we apply Elaborated action design research (e-ADR) method (Mullarkey and Hevner 2019), which serves as an extension and alteration for the original ADR (Sein et al. 2011) and provides a more structured, well-defined process map for research project management. The project follows four-cycle model in accordance with the e-ADR literature (Mullarkey and Hevner 2019). E-ADR model was proposed as an elaboration of the original ADR model for application within immersive industry-based projects. It allows the ADR team to choose the research entry point based on the current state of the problem development and specific goal of the project. Moreover, it offers an opportunity for a more structured approach in conducting an ADR by introducing Diagnosis, Design, Implementation and Evolution project phases. According to e-ADR method, each of the project cycles should by itself involve Problem Formulation, Artifact Creation, Evaluation, Reflection and Learning process activities. E-ADR also proposed the additional 8th Principle of Abstraction, which posits that every e-ADR project stage will introduce an artefact at the appropriate level of abstraction in relation with the cycle goals and activities. Altogether the e-ADR process model alterations serve as an important contribution to the overall ADR method completeness in execution and communication.

Despite the novelty of the e-ADR process model approach, it has already been applied within a number of successful doctoral and industry projects (Nunamaker et al., 2015), which have demonstrated validity and effectiveness of the method.

3.3 Empirical setting

The data is collected within the context of the Japanese branch of a multinational corporation in Tokyo, Japan. The case firm is a globally operating company with the number of associates worldwide exceeding 400.000 and providing a wide range of products and services mainly in automotive and technology fields. The case company has established its global presence in various business areas it currently operates in, such as industrial technology, mobility solutions, energy and building technology and consumer goods. It currently employs more than 5000 associates in Japan within its various locations in the country. As a major engineering and technology company, the case company aims to provide innovative solutions for its customers around the world and to improve the quality of life of people in general.

The case company is currently actively involved in developing a variety of AI-based products and has recently published code of ethics - company guidelines for the use of artificial intelligence in open access to the public. The company aims for all the products it provides to either contain AI or for them have been developed with its help by 2025. The general maxim for the case company's AI code of ethics posits that ultimately human should be an arbiter of any AI-based decisions.

The biggest business sector for the case company is mobility solutions (both hardware and software), which accounted for more than half of the total sales. One of the current strategic priorities for the company is strengthening its positioning and making the most out of its technological expertise in fields related to automated driving, connected services and mobility.

Researcher collected the data through participating in an internship with one of the departments in the case company Japan headquarters over a period of one year (June 2020 - until June 2021). The central idea behind chosen research methodology (ADR) posits that an artefact should emerge from the process of interaction with an organizational context, even though the initial research intent is guided by the researcher (Sein et al., 2011), justifying the need for a chosen empirical setting for data collection. Researcher guided the overall research process by establishing an ADR team within the company department, which consisted of 8 people (1 researcher and 7 practitioners). Moreover, due to iterative nature of ADR additional stakeholders from other departments were added depending on the research cycle. ADR team involved associates from both the engineering side (4 people) and business strategy side (4 people), allowing for sharing different perspectives and opinions.

In line with the ADR methodology (Mullarkey & Hevner, 2019; Sein et al., 2011), the study is divided into 4 cycles: Diagnosis, Design, Implementation and Evolution.

3.4 Data collection and analysis

Qualitative research methods were chosen to provide each of the ADR cycles with data collection activities. Due to practice inspired and participatory nature of ADR, qualitative methods were preferred, allowing researcher to address the problem in the real business context through a series of collaborative approaches. Qualitative data collection methods involved in-depth semi-structured interviews, participatory workshops, observation notes, analytical memos and document analysis. Data collection activities and related tasks are outlined in Work Breakdown Structure (WBS) and data collection segments in the Appendix part of the dissertation.

The data was analyzed by applying grounded theory and coding the data using the QDA tool NVivo. Researcher has followed the general guideline for conducting in-depth interviews (Boyce & Neale, 2006) in order to efficiently capture participants' perspectives, impressions and thoughts and explore the problem domain in depth. For the Diagnosis stage, semi-structured approach for the interviews was chosen and researcher has prepared a template with the questions in advance (15 questions in total), while the overall interview structure changed depending on the interviewee responses, context and general flow of the interview.

Researcher applied grounded theory methods for analyzing the interview data and referred to the relevant literature and guidelines in order to perform the analysis (Charmaz & Belgrave, 2012; B. Glaser & Strauss, 1967; Strauss & Corbin, 1990). Due to interactive and immersive nature and researcher-practitioner collaboration of an ADR project, approach to grounded theory application in a current study builds upon constructivist methods, realized through an assumption that multiple realities exist; researcher and participants may co-affect each other and the data may reflect researcher and participants' mutual constructions (Charmaz & Belgrave, 2012). The resulting portrayal of the studied phenomena takes upon an interpretive approach. Analytical memo writing was applied as a linking step between interview data coding and draft paper writing. Researcher used two memo-writing repositories; the first one went in parallel with the interview data analysis process in QDA software and was stored in the same software. The second repository was used for memo writing on the spot within the practical context to efficiently capture immediate emerging ideas based on the relevant observations. The memos were subsequently linked together in order to provide the integrated analytical view. During Implementation ADR project stage, a series of participative workshops were performed in order to support instantiation activity of the proposed artefact. We received recording permission in order to ensure continuous access to the data and subsequently analyzed it using NVivo qualitative data analysis tool.

3.5 Ethical considerations

Ethical considerations have emerged as one of the important topics within Information Systems research in the last years. Particularly, Myers and Venable (2014) proposed a set of ethical principles for design science research in IS and started a debate regarding ethics and the role of public interest in DSR. Six ethical principles have been formulated, including the public interest, informed consent, privacy, honesty and accuracy, property and quality of the artefact (Myers & Venable, 2014). According to the study, even though the nature of the ethical principles is tentative, they are aimed to serve as a basis which may be built upon, extended and, most importantly, can assist researchers in starting a dialog on the subject of ethics. Potential application of the proposed principles in ADR and related ethical implications are also communicated. We would like to address ethical considerations within the scope of the study in the following part.

We ensured the privacy of stakeholders (developers, engineers, business side managers) through anonymization of related information, such as names and title roles. No internal information has been disclosed to the third parties and we strictly observed inquiry process in regard to recording data (voice records for the interviews and participatory workshops) to ensure continuous access needed for the data analysis. All the participants have been informed of data collection in advance and we have been granted permission to conduct research activities. Subsequently, interpretation of data was presented to the stakeholders following artefact instantiation and finalizing the reflection part of our project.

Chapter 4. Action Design Research project phases

Similar to the case described in Mullarkey & Hevner (2019), current study is realized as an immersive practice-based project dealing with a particular class of problems (algorithmic accountability), in which the researcher is facing a challenge of no prior existing artefact to address the problem. Researcher applies the elaborated Action Design Research method presented in Mullarkey & Hevner (2019), which serves as an extension and alteration for the original ADR article by Sein et al. (2011). Mullarkey & Hevner (2019) propose four stages of ADR project realization, namely Diagnosis, Design, Implementation and Evolution.

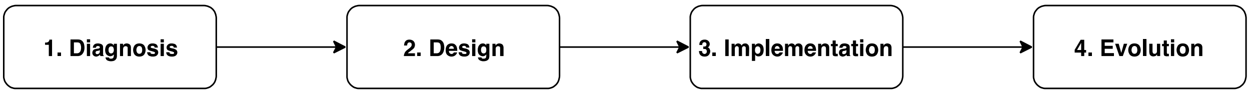


Figure 5. Elaborated Action Design Research cycles summary adapted from Mullarkey and Hevner (2019)

The following chapter will discuss in detail all the related activities realized in each of the phases resulting through the appropriation of Action Design Research method in the current study in order to visualize the path of progress and emerging insights.

4.1 Diagnosis

Mullarkey and Hevner (2019) propose Diagnosis stage as a necessary initial step prior to designing a new ensemble artefact. The main goal of the Diagnosis stage is to instigate researcher-practitioner intervention and reach mutual understanding by thorough investigation and definition of problem domain and its importance, as well as to evaluate an IT solution class (Mullarkey & Hevner, 2019). Main activities of the Diagnosis stage include identification of relevant kernel theories, specifying overall goals of the ADR project and related socio-technical artefacts (Mullarkey & Hevner, 2015, 2019).

e-ADR model also posits that every research project phase by itself has to go through Problem Formulation, Artefact Creation, Evaluation and Reflection and Learning phases (Mullarkey & Hevner, 2019). This division of the whole research project into smaller chunks allows for more efficient implementation of related research activities and ensures better project management.

The two important learning areas during the Diagnosis stage include the full understanding of the domain for application and awareness of the related knowledge base (Mullarkey & Hevner, 2019). Application domain understanding refers to researcher fully grasping opportunities, strengths, weaknesses and constraints specifically related to the organization where the research project is conducted. Knowledge base includes the related study fields which will subsequently inform the design of the artefact (Mullarkey & Hevner, 2019).

According to Mullarkey & Hevner (2019), an artefact resulting through performing an ADR Diagnosis stage may vary from conceptualization of the problem domain to specification of requirement definitions. In the current study the resulting artefact of the first ADR project stage is conceptualization of the problem domain of the case organization. This conceptualization is realized through performing a set of data collection activities within the organization as well as review of relevant literature to inform the knowledge base component of the Diagnosis stage. The timeline and detailed description of activities performed at the Diagnosis stage of the project are outlined below.

During the Diagnosis stage research problem was discussed and evaluated during the meetings with the ADR team practitioner side representatives (business strategy side managers), conducted both face-to-face and online. The ADR project idea was mutually explored by both the industry-side associates and the researcher, drawing upon Principle 1 (Practice-inspired Research) ad Principle 3 (Reciprocal Sharing) of the Design Science methods.

ADR literature does not prescribe particular data collection and analysis methods to be used during each of the project stages and researcher is free to interpret and define the most appropriate methods on her own (Sein et al., 2011). Researcher performed a set of five semi-structured qualitative interviews in order to gain understanding of the problem domain, identify current practices, tools and methods through which algorithmic accountability is realized in the case company and concerns (if any) of relevant stakeholders and accountability implementation in developing algorithmic systems. The data was further analyzed by applying grounded theory and coding the data using the QDA tool NVivo. Researcher has followed the general guideline for conducting in-depth interviews (Boyce & Neale, 2006) in order to efficiently capture participants' perspectives, impressions and thoughts and explore the problem domain in depth. Due to the chosen semi-structured approach for the interviews, researcher has prepared a template with the questions in advance (15 questions in total), but the overall interview structure changed depending on the interviewee responses, context and general flow of the interview.

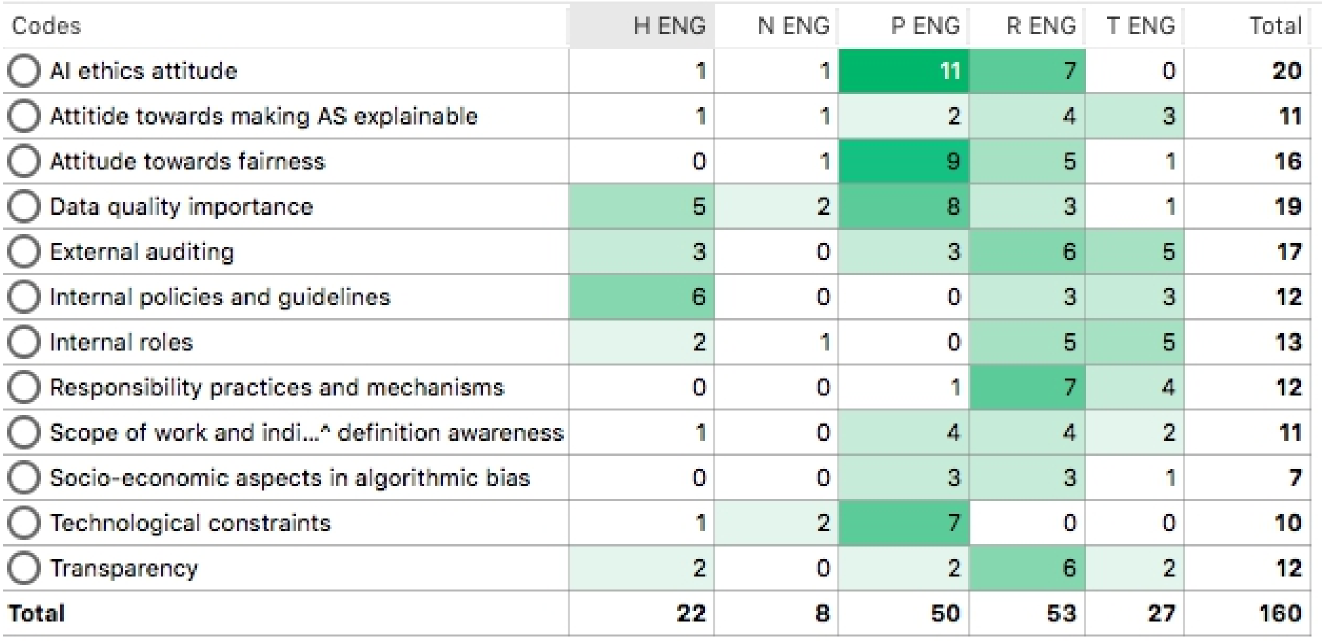
All of the interviews were performed face to face in the case company headquarters in Tokyo and 4 out of 5 interviews were recorded with the prior permission received by the researcher. Participants were provided with the disclaimer information explaining the objectives, background information and data usage conditions. Participants were then asked to answer questions related to their current scope of work, personal opinions about algorithmic bias and fairness issues, algorithmic accountability practices realized in the case company, tools and methods presently applied for software risk mitigation, software development process details and transparency in sharing the data with the customers and end users; impressions relating to AI ethics, algorithmic bias; awareness of ethics-related internal documentation and guidelines; necessity of external auditors to check for compliance for algorithmic accountability and internal roles for ensuring responsibility. The interview data was further transcribed and analyzed using QDA software.

Researcher applied grounded theory methods for analyzing the interview data and referred to the relevant literature and guidelines in order to perform the analysis (Charmaz & Belgrave, 2012; B. Glaser & Strauss, 1967; Strauss & Corbin, 1990). In-depth qualitative interviewing fits grounded theory methods especially well (Charmaz & Belgrave, 2012) and provides the researcher with "an open-ended, in-depth exploration of an aspect of life about which an interviewee has substantial experience" (Charmaz & Belgrave, 2012). Since an interview is a flexible data collection technique, new ideas may emerge during the process and the researcher may pursue these emergent issues and leads. Due to interactive and immersive nature and researcher-practitioner collaboration of an ADR project, approach to grounded theory application in a current study builds upon constructivist methods, realized through an assumption that multiple realities exist; researcher and participants may co-affect each other and the data may reflect researcher and participants' mutual constructions (Charmaz & Belgrave, 2012). The resulting portrayal of the studied phenomena takes upon an interpretive approach.

The interviews were transcribed and coded in a three-step process using the QDA software tool NVivo. Grounded theory methods prescribe that the interview coding process includes at least two stages, through assigning the initial codes (also referred to as an open coding) to the data and instigating the analytic decision-making process and subsequently perform the selective coding (focused coding), which assists the researcher in further synthesizing, sorting and conceptualizing the data (Charmaz & Belgrave, 2012). The initial open coding resulted in 36 codes being assigned to the data through line-by-line coding approach and by using active terms in order to describe and define the data. In the selective coding phase, researcher has analyzed the most frequently reappearing codes from the open coding stage and further synthesized the data by assigning more precise and general focused codes in order to lay the foundation for the next stage in the analytical process. The third coding layer includes the five guiding Principles for Accountable Algorithms and a Social Impact Statement for Algorithms (Diakopoulos et al., 2018). The resulting coding framework including the three coding layers is presented in Table 1.

Researcher applied analytical memo writing as a linking step between interview data coding and draft paper writing. Grounded theory researchers suggest that memo writing serves as an essential intermediary stage for instigating analytical process (Bryant & Charmaz, 2012; Charmaz, 2006; Charmaz & Belgrave, 2012; B. G. Glaser & Holton, 2007). Memo writing is a process of writing theoretical notes about the data and the related emerging conceptual linkages. Writing analytical memos in grounded theory is known to be a fundamental process of researcher engagement with the data and transforming it into the theory: "The writing of theoretical memos is the core stage in the process of generating grounded theory. If the researcher skips this stage by going directly to sorting or writing up, after coding, she is not doing grounded theory" (Bryant & Charmaz, 2012). Memo writing also guides the researcher in the subsequent research activities, including data collection, analysis and coding. Memo writing allows the researcher to stop and think about the data, elaborate the specific processes, initiate new ideas and define the gaps (Charmaz & Belgrave, 2012). Once synthesized and sorted, memos may serve as a foundation for the formulation and presentation of theory (Bryant & Charmaz, 2012). In the current study analytical memos served as a basis for the overall draft paper outline.

Researcher used two memo-writing repositories, the first one went in parallel with the interview data analysis process in QDA software and was stored in the same software. The second repository was used for memo writing on the spot within the practical context to efficiently capture immediate emerging ideas based on the relevant observations. The memos were subsequently linked together in order to provide the integrated analytical view.



Codes	H ENG	N ENG	P ENG	R ENG	T ENG	Total
AI ethics attitude	1	1	11	7	0	**20**
Attitide towards making AS explainable	1	1	2	4	3	**11**
Attitude towards fairness	0	1	9	5	1	**16**
Data quality importance	5	2	8	3	1	**19**
External auditing	3	0	3	6	5	**17**
Internal policies and guidelines	6	0	0	3	3	**12**
Internal roles	2	1	0	5	5	**13**
Responsibility practices and mechanisms	0	0	1	7	4	**12**
Scope of work and indi...^ definition awareness	1	0	4	4	2	**11**
Socio-economic aspects in algorithmic bias	0	0	3	3	1	**7**
Technological constraints	1	2	7	0	0	**10**
Transparency	2	0	2	6	2	**12**
Total	**22**	**8**	**50**	**53**	**27**	**160**

Figure 6. Coding crosstab query

The coding crosstab query in Figure 6 represents the general pattern for codes frequency in the interview data. In total 160 codes were assigned across the interviews. For the open coding step researcher applied line-by-line coding in order to make the initial sense of the data and discover participants' views. In accordance with the grounded theory guidelines for analyzing data by Charman & Belgrave (2012), the codes assigned were action codes, reflecting action in the data and assisting in keeping the analysis more specific (e.g., "Claiming that making AS explainable is unnecessary", "Expressing skepticism regarding AI ethics", "Linking ethics with aspects outside of the technical area"). In the next stage of analysis, the initial open codes were grouped together based on common linkages and reappearing patterns and themes. This process resulted in assigning 12 focused codes, serving as the outline for further analytical work and connecting focused codes with the Principles for Accountable Algorithms and a Social Impact Statement for Algorithms (Diakopoulos et al., 2018). In Figure 6 the commonly reappearing coding references in the interview data are visualized through the heatmap, where the biggest number of coding references across the whole structure are highlighted with the darker green colors. As seen in the Figure 6, focused codes with the greatest number of coding references are as follows: AI ethics attitude, Data quality importance, External auditing and Attitude towards fairness.

The resulting analytical framework served as a basis for introducing empirical claims. Empirical claims were proposed by the researcher as the concluding step and key artefact in the Diagnosis phase. Empirical claims arise from analyzing the interview data and may be referred to as a set of identified insights, issues and patterns derived from the data. The outline of the empirical claims for the Diagnosis stage of the study is presented in Table 7.

Empirical claim	Focused coding	ACM coding	Number of coding references
EC1 Participants tend to be skeptical towards AI ethics in general; AI perceived as an area of cutting-edge research in specific departments within the organization, not something dealt with daily	AI ethics attitude	Fairness	20
EC2 Challenges related to providing explanations for AS outweigh the benefits	Attitude towards making AS explainable	Explainability	11
EC3 Limited understanding of how fairness can be introduced through AS design	Attitude towards fairness	Fairness	16
EC4 Participants recognize that the issues relating to data quality assurance should receive more attention, datasets is a primary source of bias	Data quality importance	Accuracy	19
EC5 External auditing is not necessary, various issues related to its implementation (e.g. corruption, AI development hindrance)	External auditing	Auditability	17
EC6 Lack of awareness about AI ethics and ethics related internal guidelines	Internal policies and guidelines	Responsibility	12
EC7 Developers should not necessarily be held accountable for algorithmic bias	Internal roles	Responsibility	13
EC8 Limited understanding regarding hierarchy in responsibility levels within the organization and personal accountability scope	Responsibility practices and mechanisms	Responsibility	12

EC9 Lack of awareness of «algorithmic accountability», «algorithmic audit» concepts	Scope of work and individual documentation & definition awareness	Responsibility	11
EC10 Civil society organizations should not necessarily be a part of the AS development process	Socio-economic aspects and algorithmic bias	Fairness	7
EC11 Constraints related to current technology play a role in algorithmic bias	Technological constraints	Accuracy	10
EC12 Participants agree on a high level of transparency for organizational processes in general	Transparency	Auditability, explainability	12

Table 7. Empirical Claims for the Diagnosis stage of the study

The summary of findings and description for the Design stage artefact (Empirical claims, EC) is discussed below.

EC1: Participants tend to be skeptical towards AI ethics in general; AI perceived as an area of cutting-edge research in specific departments within the organization, not something dealt with daily

EC1 was derived from the data based on the largest number of coding references (20) across the interviews. The data reflected overall skeptical perception of AI-related ethics within the engineers. One participant referred to algorithmic systems used in hiring for top IT companies to emphasize the primary objective of using such systems (efficiency) versus developers taking into account potential ethical issues resulting from using such systems:

“The people making the algorithms, they are less concerned about ethics and more concerned about getting the best person for the job… “ (P ENG)

Another participant highlights the possibility of AI ethics being a hindrance factor in AI technology development resulting from enforcing ethics:

“I definitely think it is very important, but at the same time, we got to be careful because it is something very hard to control…Development of AI in general, like any kind of big system is very organic and if you just try to hold it down and keep it on a leash to make sure it is always

ethical, «perfectly» ethical, I feel like it is going to push it back. To a certain extent, pushing the boundaries of it is important, but having too much ethics to take into account might harm it, hold the development down …" (R ENG)

EC2: Challenges related to providing explanations for AS outweigh the benefits

EC2 was based on 11 coding references, with interviewees highlighting the related issues and challenges of making algorithmic systems explainable. Participants generally admit the good behind making the decision-relevant aspects of using algorithms visible and understandable, but also point out that at the current stage the disadvantages and efforts needed to realize this outweigh the benefits:

*"You can provide transparency, but I don't think that it would be useful, because it is too complex… "(*T ENG)

Moreover, one interviewee points out the importance of majority of the general population being technologically illiterate compared with the people involved in the design and development of the algorithmic systems:

*"Yes, it would be good to let consumers know what is going on, but most of the time they would not care. And if they do care, most likely they are illiterate in a technological sense. It is confusing, because if you make it transparent to the public, the public won't understand anything… "(*R ENG)

EC3: Limited understanding of how fairness can be introduced through AS design

Participants tend to have a limited understanding on how fairness can be introduced in the algorithm at the design stage. In general interviewees refer to AS as a type of technical system which initially is not fit to consider the whole array of issues related to potential just/unjust treatment of the end users:

"We do not mean algorithms like that. We do not make exceptions for race, gender, stuff like that ". (P ENG)

Secondly, participants rejected the idea of importance for agenda specification in order to identify and eliminate the potential issues and sensitivities at the pre-operational phase. Moreover, one interviewee reflects on the possible issue of preferential treatment and discrimination resulting from the actual attempts to make AS fair:

"If we make an algorithm that preferentially treats a race or gender, it would be discriminating against other races…". (P ENG)

EC4: Participants recognize that the issues relating to data quality assurance should receive more attention, datasets is a primary source of bias

EC4 elicited from 19 coding references corresponding with “Data quality importance” focused code and “Accuracy” ACM code. This is the second largest coding reference group, with interviewees stressing the importance of the quality for the initial datasets that the algorithms are fed with:

“Let's say, I created an algorithm for the “M” company. I know that all the other companies surrounding this company are not hiring Asians. The data that I use, most of the Asians won't be hired…and I feed that data to my algorithm…I would say, get a better set of data, a dataset. Get more unbiased dataset. Unless you are deliberately writing algorithms that discriminate, which itself would be illegal…” (P ENG)

“If AI input is limited, then the output also somehow we can define… If we want to use one function, then from the input and output relationship the whole system can be designed…” (H ENG)

EC5: External auditing is not necessary, various issues related to its implementation (e.g. corruption, AI development hindrance)

Participant views regarding the possibility of introducing a third-party auditor in order to inspect and ensure compliance for accountable algorithmic practices realized within organizations in general are mixed. Some interviewees point out the array of possible issues related to politics, corruption and profit distribution:

“I feel like it should be definitely a government agency, removed from bias and profits. Because it is a public good, we all want technology and all want efficient AI, but it should be completely dissociated with politics, we want people who are technically minded to do this kind of stuff…” (R ENG)

Another participant suggests that the third-party auditor involvement should only be necessary in case of apparent law violation:

“I think that it is not necessary, however, that laws should be implemented and if a company breaks one of these laws, then a third party could investigate…” (T ENG)

EC6: Lack of awareness about AI ethics and ethics related internal guidelines

EC6 was based on 12 coding references from the interview data, reflecting participants' AI ethics awareness and awareness of related internal guidelines available within the case company. All the participants failed to recall the major AI ethics-related guideline published by the case company in the beginning of the same year. The company code of ethics for AI served as a significant milestone as one of the first guidelines of this kind for a leading multinational technology enterprise in securing consumer trust and serving as an important facilitator for organization's competitive and brand strategy. The aforementioned code of ethics for AI also generated major publicity and media coverage in the beginning of the year. Moreover, when asked about internal ethics-related guidelines, participants did not provide many references and knowledge related to the topic. However, one senior employee with a significant prior experience in the automotive field referred to industry standards such as Automotive Spice and CMM.

EC7: Developers should not necessarily be held accountable for algorithmic bias

EC7 is based on 13 coding references in the interview data, which reflect participants' views on developers being accountable in case of an algorithmic bias. One participant suggested that the main responsibility should lie with the actor releasing the software in accordance with the business contract:

"If someone approves the software and releases it organization-wise, then the representative should take responsibility for this problem…Based on the business contract, who releases the software and gets money from this act…" (H ENG)

Some respondents were hesitant to provide the definite answer, but suggest that developers should not necessarily be held accountable depending on a particular situation and context:

"I guess it depends a lot on the case, but developers are not necessarily to be held accountable…I would not account them necessarily for everything that happens. I honestly cannot say who should be held accountable…" (R ENG)

The answers reflect overall conflicting nature of the issues related to algorithmic accountability, with participants struggling to define the boundaries and extent of developers being held accountable in case of discovered bias.

EC8: Limited understanding regarding responsibility distribution within the organization and personal accountability scope

Coding references in this category (12 in total) indicate that participants' comprehension and awareness of how responsibility is distributed between the stakeholders within the case company is limited. Moreover, the scope of one's own accountability is vague. In addition, the data revealed that despite generally high level in transparency both regarding the end users, partners and individual actions being tracked down (as reflected in EC12), some respondents tend to fully shift responsibility onto supervising associates.

EC9: Lack of awareness of «algorithmic accountability», «algorithmic audit» concepts

EC9 is derived from 11 coding references across the interview data, with participants showcasing limited understanding of the concepts in question. Particularly, one respondent has a following perception of the term "algorithmic accountability" as an algorithm dealing with the human damage potential as opposed to its actual meaning:

"I have never heard about it, but when you say «algorithmic accountability», I think you are just saying…having some sort of an algorithm that deals with any kind of human damage potential, just in a way to minimize the damage and how the developers have the responsibility to do that…" (R ENG)

EC10: Civil society organizations should not necessarily be a part of the AS development process

EC4 elicited from 7 coding references corresponding with "Socio-economic aspects and algorithmic bias" focused code and "Fairness" ACM coding layer from the interview data. The data revealed that participants do not consider the involvement of civil society organizations in the design and development stage of algorithms as a necessary measure. In particular, one respondent suggested that revealing development-relevant information may become an issue of information disclosure and trade secrecy:

"Discussion yes, design no. Algorithm by itself is a company property. Unless these are open-source algorithms, which there are, company has every right to preserve it as a secret…" (P ENG)

EC11: Constraints related to current technology play a role in algorithmic bias

EC11 is derived from 11 coding references in the data referring to various technology-related constraints as one of the facilitating factors in algorithmic bias. One participant recalled an example of an algorithmic system used in the detection of faces (face-detection algorithm) by one of the multinational IT companies not working properly for Black people due to one of such constraints and receiving major negative publicity due to racism connotations:

"The only reason why it wasn't detecting Black people is because of the way we detect faces using structures, and it depends on contrast, light of the image. With Black people, what happened was there was a washed-out contrast. So, when you put something like an age detection algorithm on them, it couldn't find ages…" (P ENG)

Participant suggested that fairness and ethics should be addressed separately depending on the context of the situation and that ethics is not necessarily violated when the primary reason for algorithmic bias is a technological constraint. Similar to the point in EC3, the idea of prior agenda consideration and investigation for field sensitivity for bias is not considered by the participants.

EC12: Participants agree on a high level of transparency for organizational processes in general

EC12 is based on 12 coding references from the interview data. Respondents acknowledge in general high transparency for the internal organizational processes, including individual actions being tracked down and documented and information disclosure for the related external parties (including partners and customers):

"I would say it is very transparent, and from what I see, from the PPT and other things, we are really open about what we do within the company, how we want to do it, how to process work. There is of course secrecy, you cannot just go around and show it to everyone, because this is the company's intellectual property, and it is worth a lot of money. But there is a fair level of transparency. I think it is fairly difficult, because when you put a lot of money in something, you usually would like to know exactly how it works. So, we are really doing a good job at explaining how the things work. All the failure savers and so on…" (R ENG)

The analytical framework discussed above (Empirical claims, EC) serves as a key artefact for the Diagnosis stage of the ADR process. Our study will proceed with the Design phase in order to create a prototype of the solution in order to solve identified conflicting issues derived from the insights based on the Empirical Claims.

4.2 Design

According to the elaborated action design research model outlined in Mullarkey & Hevner (2019), design stage follows the diagnosis part of the study. In the Diagnosis stage thorough investigation of the problem domain serves as a focus and the researcher-practitioner collaboration is involved in evaluating and identifying relevant socio-technical artefacts, kernel theories and reaching a mutual understanding in terms of the goals of an ADR project (Mullarkey & Hevner, 2019). Our study has completed the Diagnosis stage by producing a set of Empirical Claims derived from the data analysis in the initial stage and serving as a key concluding artefact for the first stage of the project.

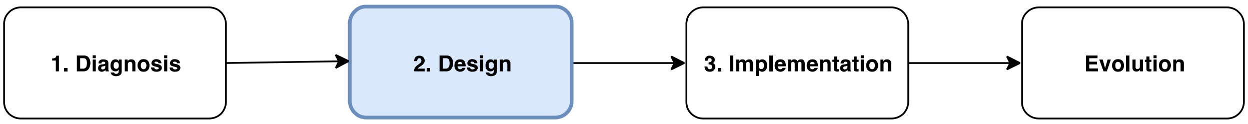


Figure 7. Elaborated Action Design Research process model cycles adapted from Mullarkey and Hevner (2019): Design stage

This chapter will provide a description for the activities realized during the Design stage of the ADR project. Elaborated ADR model posits that the Design stage "provides a set of activities over the search space of possible design candidates" (Mullarkey & Hevner, 2019, p. 10). Design stage of the ADR project will address the problem identified during the preceding Diagnosis stage and will allow the ADR team to move towards Implementation phase. During the Design stage the ADR team will contribute to the design of innovative ideas to solve the given problem in order to contribute to both the practical and theory streams of knowledge. According to Mullarkey & Hevner (2019), some of the example artefacts for the Design stage of an e-ADR project are methods, models, design principles, design features or architectures. Sein et al. (2011) suggests that Design activities should be incorporated within the Building, Intervention and Evaluation (BIE) phase, whereas Mullarkey & Hevner (2019) argue that clear separation of design activities in the proposed Design stage is necessary. As this study follows the e-ADR model structure, researcher explicitly separates the design activities in the Design stage.

Key learning outcomes and implications from the Diagnosis stage of this study will be addressed in order to inform the design features of the proposed system. Current ADR literature and studies applying ADR methods (including the original and e-ADR models) approach the design activities differently. Some researchers undertake UX design and agile approaches (Keijzer-Broers & de Reuver, 2016), whereas others initiate prototype building inspired by relevant literature (Haj-Bolouri, 2019), a set of organizational and technological interventions in order to produce relevant design principles for the problem domain of competence management systems (Niemi & Laine, 2016), or assessing affordances for the wildlife management analytics system (Pan et al., 2020).

Insights derived from the Diagnosis part of the study clearly show that the participant engineers tend to focus on the efficiency side of the algorithms, neglecting the potential ethics implications and biases from their realization. The data reflects that even though bias and unfairness are recognized as a problem space in the algorithmic domain, it is something that exists outside of the practical scope, relatively less important than maximizing the efficiency of an algorithmic system itself and very challenging to implement. However, ethical consequences of using algorithms are not necessarily pre-fixed in the design of the algorithmic systems and organizations should be mindful of indirect biases (Martin, 2019).

Secondly, EC data reflects the current state of ambiguity and lack of awareness regarding how responsibility is distributed between the stakeholders within the case company. However, the design of the algorithms calls for clear understanding of responsibilities and roles of the decision system (Martin, 2019). It is therefore necessary for the employees to have a better awareness of one's responsibility scope and responsibilities of their supervisors and other associates, as the lack of such understanding may lead to the culture of blame shifting as reflected in the EC8.

As our study aimed to reflect the socio-technical nature of algorithmic accountability as a concept, it is essential to highlight the importance of participant engineers' views on algorithmic neutrality. According to the insights derived in the previous phase of the study, engineers perceive algorithms as systems possessing face value objectivity, a sequence of computational steps designed to produce the output based on the input and maximizing the efficiency in solving a problem it is intended to solve. For example, as one of the participants has noted, to make the algorithmic system produce unfair results, developers would need to "deliberately sabotage it". However, one of the assumptions in the current study and the argument reflected in the vast scope of relevant literature (Katell et al., 2020; Martin, 2019; Mohseni et al., 2018; Polack, 2020; Wieringa, 2020) indicates that the individuals involved in the design and development of the algorithmic system cannot simply be separated from its decisions, as the bias can find its way into the algorithms due to the many ways these individuals stay involved in the algorithmic decisions. If AI/AS engage in human lives and communities as quasi-autonomous agents, then they must be expected to follow the community's social and moral norms (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). The narrative of algorithms free of bias by default is a misleading one and appropriate measures should be taken to achieve its obliteration. In order to ensure that the algorithms in organizations are utilized responsibly, stakeholders need to consider the notion of value-laden algorithms. It is therefore essential to introduce an element that enforces associates to reconsider the narrative of neutral algorithms in order to facilitate the culture of responsibility and its acceptance.

Moreover, EC data reflects the current misalignment between the state of research in the field of ethically aligned AI systems, including algorithmic accountability, and the practical state of participant engineers' awareness of ethics. The Diagnosis phase data revealed that participants have limited understanding of how fairness can be introduced into algorithmic design, including awareness of pre-development sensitivity of specific algorithm application agenda, as witnessed in cases similar to COMPAS recidivism algorithm (Larson et al., 2016; Tan et al., 2018; Washington, 2019). Additionally, even though the case company intends to introduce AI in all its products by 2025, some participant engineers tend to refer to AI as an area of higher research and complexity, which is outside of the scope of their daily duties and responsibilities. The aforementioned issues should be considered by organizations interested in establishing the culture of responsibility internally, as the data revealed that simply introducing ethics-encouraging prescriptive-based guidelines does not necessarily lead to tangible results and systemic awareness.

Finally, accountability systems used within the organizations are a product of business ethics and corporate strategy (Binns, 2018). In this sense algorithms utilized in private firms would differ from public and non-profit organizations, whereas in the former case companies would have an incentive to nudge customers in the right direction in order to maximize potential profits. Interview data from the Diagnosis part of study reflects how engineers tend to view responsibility from pragmatic and financially oriented viewpoint, linking organizational ethical considerations to purely compliance-related incentives. Participants argue that maximizing revenue and reputational concerns are the primary objectives for the company to stay compliant, while ethical considerations will be an area of secondary importance. This discussion is out of scope of the current study; however, future research may address the potential linkage between company brand image, reputation and efficiency in implementation of ethically aligned AI systems.

Guided by our research objective, we set out to develop a set of design principles that would improve algorithmic accountability within the case company context. In constructing these design principles, we relied not only on the Empirical Claims formulated in the Diagnosis part of our study, but also on the principles for the ethical and values-based design, development and implementation of autonomous and intelligent systems established by IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). As a seminal document in the field of ethically aligned design, it was created by more than 700 researchers and experts, addressing a wide range of issues related to instantiation of value-laden AI/AS and human-centric design and aims to provide guidance for a wide range of stakeholders and audiences. The general principles of Ethically Aligned Design consider both the role of the AI/AS creator, operators and any other affected parties or stakeholders (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). Accountability serves as one of the fundamental principles for ethically aligned AS and lack of accountability is acknowledged to present a major challenge for the implementation of the AS development and design. Another document that was utilized during the process of constructing the design principles was a Statement on Algorithmic Accountability and Transparency published by the Association for Computing Machinery (ACM) US Public Policy Council (Association for Computing Machinery US Public Policy Council (USACM), 2017). This statement is consistent with the ACM Code of Ethics and Professional Conduct (Anderson, 1992) and was developed to support the benefits of algorithmic decision-making while addressing a wide range of concerns related to the impact of algorithms on society.

Design principle	**Explanations and rationale**

The principle of raising awareness of ethics and ethical literacy	Limited understanding of ethical implications of using algorithms and misalignment between the state of research in the field of ethically aligned AI systems (including industry-produced Codes of ethics) and the practical state of participant engineers' awareness of ethics. It is necessary to educate stakeholders on societal impacts of designing, developing and implementing AS.
The principle of value-based design incentivization and appreciation of AS deployment context	Values-based design methods should be put in the center of the technical system development in order to create sustainable systems providing not only economic value to the organizations but increasing human and societal well-being. In order to ensure that the algorithms in organizations are utilized responsibly, stakeholders need to consider the notion of value-laden algorithms, as opposed to free of bias, neutral narrative of algorithm usage.
The principle of actionable guidelines	Industry guidelines for AI/AS ethics should include actionable statements rather than descriptive principle and value-related formulations, which are too vague to translate into tangible results. Incorporating ethics into technology design agenda also deals with translating the norms into language accessible to different levels of stakeholders (e.g., policy nuances into technical context).
The principle of transparency	Lack of transparency increases the difficulty in achieving accountability (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). Operation of AS should be made transparent to a wide range of stakeholders, however transparency (also addresses explainability and traceability) may need to be targeted towards specific type of decision and purpose (Ananny & Crawford, 2018).
The principle of stakeholder responsibility clarification	Design of algorithms calls for clear understanding of responsibilities and roles of the decision system (Martin, 2019). Clarifying participant dynamics helps to ensure more transparent provision of information and improved interpretation of the system usage context.

Table 8. Algorithmic accountability canvas design principles and corresponding explanations and rationale

In order to visually represent the proposed integrated Algorithmic Accountability Canvas tool, we decided to adopt ‘‘business model canvas’’ (BMC) approach developed by Osterwalder & Pigneur (2010) as a method for supporting the design of business models. The concept of BMC was initially introduced in a doctoral dissertation “The Business Model Ontology: A Proposition in a Design Science Approach” by Alexander Osterwalder (A. Osterwalder, 2004). Research goal of the dissertation was to address the concept of business models in order to provide the basis for new management tools in information systems and strategy, possibly software based, allowing the firms to express their business logic (A. Osterwalder, 2004). The business model ontology prototype and its instantiation developed by A. Osterwalder resulted in a tool which is aimed at “facilitating the description of a business model” (A. Osterwalder, 2004, p.3). The initial tool was developed through application of design science research approach, where researcher was involved in building and evaluating an artefact named BM^{2}L, which served as a basis for a business model ontology (A. Osterwalder, 2004).

Business model ontology has since received much attention both in the academia and on the practitioner side, having received widespread adoption by businesses not only for designing, describing, visualizing and assessing the current state of business models, but also for future business innovation and as a lean startup template (Fritscher & Pigneur, 2014). In academia it was previously addressed to analyze big data applications (Muhtaroglu et al., 2013), to adopt service logic in business model thinking by introducing Service Logic Business Model Canvas (Ojasalo & Ojasalo, 2018), to facilitate a system development for effective budgeting (Dudin et al., 2015), to improve investment processes (Sort & Nielsen, 2018) and others. Moreover, some adaptations of the original Business Canvas Model (Alexander Osterwalder & Pigneur, 2010b) have been introduced, including Value Proposition Designer model, mostly focused on the customer side (Alexander Osterwalder et al., 2014) and Lean Canvas model, which is predominantly aimed for entrepreneurial use in the startup field (Maurya, 2012, 2014). However, the original Business Model Canvas remains the most widely accepted both on the practitioner side for the existing and new businesses and in academia alike.

Researcher has adopted the original Business Model Canvas model (Alexander Osterwalder & Pigneur, 2010) as a method to support the design of the proposed Algorithmic Accountability Canvas tool through application of previously outlined Design Principles. As the current study deals with the development and instantiation of an artefact in a business context through researcher’s continuous involvement in case company’s processes and interaction with a variety of stakeholders within the organization, the adaptation of the chosen Business Model Canvas was deemed as the most efficient method to communicate the conceptualization of the proposed tool in a visual way, which is easily understandable, practical and can be directly translated into the business language. Moreover, both the managers from the business strategy, sales and marketing, as well as engineering side in the case company were already acquainted with the Business Model Canvas itself, as the company has a history of implementing it for new business development and ideation activities. Lastly, as the Business Model Canvas has been proven to be an efficient tool in engaging stakeholders ranging from executives in large companies to SMEs and entrepreneurs by serving as an interactive foundation, it is well-aligned with the Principle of actionable guidelines outlined in the Design Principles introduced earlier in this chapter.

Alexander Osterwalder & Pigneur (2010) BMC	**Algorithmic Accountability Canvas Mapping**	**Design Principles**
Key partners	Key actors (KAC)	
Key activities	Key activities (KAT)	The principle of raising awareness of ethics and ethical literacy The principle of value-based design incentivization and appreciation of AS deployment context The principle of actionable guidelines
Key resources	Key resources (KR)	The principle of raising awareness of ethics and ethical literacy
Value proposition	Value proposition (VP)	
Customer relationships	Stakeholder responsibility clarification (SRC)	The principle of stakeholder responsibility clarification
Customer segments	Excluded from AAC	
Channels	Transparency (explainability, traceability) (TSP)	The principle of transparency
Cost structure	Cost structure/budget (CS)	
Revenue streams	Value created (VC)	
	*Data (DT)	The principle of value-based design incentivization and appreciation of AS deployment context The principle of transparency
	*Evaluation and monitoring (internal audit) (IA)	The principle of transparency
	*Independent oversight (external audit) (EA)	The principle of transparency

Table 9. Algorithmic Accountability Canvas development based on corresponding BMC elements (Osterwalder & Pigneur, 2010), including mapping to Design Principles

Algorithmic Accountability Canvas development mapping is outlined in Table 9. Researcher has excluded Customer segments from Algorithmic Accountability Canvas due to primary goal of the artefact to assist organizations in improving algorithmic accountability as an internal system; therefore, customer segmentation remains an area of secondary importance for the proposed tool. Moreover, three elements were added to the original BMC (Alexander Osterwalder & Pigneur, 2010), namely Data, Evaluation and monitoring (internal audit) and Independent oversight (external audit). In Algorithmic Accountability Canvas, Data refers to the proposed set of tools aimed to facilitate reflection, documentation and communication of datasets and models utilized within the organization. Evaluation and monitoring (internal audit) propose an internal audit framework structure for algorithmic auditing, whereas Independent oversight (external audit) serves as an external regulatory body to ensure compliance, which depends on the current legislation in particular state.

Key actors (KAC)

Key actors (KAC) section of Algorithmic Accountability Canvas is aligned with Key partners part of the Business Model Canvas (Alexander Osterwalder & Pigneur, 2010). As improving algorithmic accountability is seen in the scope of the current study as an internal organizational issue which needs to be solved, partnerships play a secondary role in the context of the proposed model and key organizational actors are addressed instead. Key actors part refers to company associates, namely all actors involved in the design, development and deployment of AS in the organization. Moreover, it is important to note that due to importance of ethical implications and indirect biases in the AS design problem space, a wide range of stakeholders should be addressed.

Key activities (KAT)

Key activities (KAT) section of Algorithmic Accountability Canvas refers to the main activities and processes the organization will need to undertake in order to improve algorithmic accountability. Key activities is subsequently divided into the following directions: Educate developers and designers of AS; Value-based incentivization and Develop supports tools.

1. Educate developers and designers of AS part relates to the following issues:

a. Recognition of societal impacts of AS as a problem space

This activity is aligned with the following Design Principles: The principle of raising awareness of ethics and ethical literacy and The principle of value-based design incentivization and appreciation of AS deployment context (Table 8). Recognition of societal impacts of AS is central for facilitating accountability, therefore organizations should make an effort to educate, engage and inform designers and developers of AS regarding the negative and positive potential outcomes and impacts of AS deployment for the society.

b. Improving an understanding of how fairness can be introduced into AS design

Findings from the Diagnosis stage of the study (EC3) revealed that participants tend to have a limited understanding on how fairness can be introduced in the algorithm at the design stage, including the general undermining of the idea of importance for agenda specification in order to identify and eliminate the potential issues and sensitivities of AS at the pre-operational phase. A wide range of stakeholders involved in the design, development and deployment of AS organization-wise should be involved in corporate training activities, including workshops, lectures and internal e-learning tools in order to advance the knowledge on fairness by design concept in AS.

c. Importance and diversity of the existing cultural norms among the users of AS

Importance and diversity of the existing cultural norms among the users of AS is one of the central themes in the field of Ethically Aligned Design (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). Stakeholders should be educated on the aforementioned topic in order to enable a cross-cultural dialogue of ethics in technology, facilitate innovation and contribute to human well-being and society on the whole.

2. Value-based design incentivization

Value-based design-incentivization activity in Key activities part refers to necessity to integrate the concept of value-laden AS as opposed to "neutral" narrative of algorithms to mitigate the associated risks, as well as to put value-based design methods in the center of the technical system development. This activity is aligned with The principle of value-based design incentivization and appreciation of AS deployment context (Table 8). Moreover, introducing actionable AI/AS ethics guidelines and translating norms into language accessible to a wide range of stakeholders should be implemented. For example, prescriptive statements (e.g., "should", "encouraged") need to be replaced with enforcing clauses instead.

3. Develop support tools

A set of tools needs to be introduced internally in order to support implementation of the proposed system. Researcher has identified the need for developing and integrating corporate training tools (including e-learning corporate training courses and materials), internal audit system proposed in Internal Audit section of Algorithmic Accountability Canvas and data reflection tools introduced in Data section.

Key resources (KR)

Key resources (KR) section relates to major resources needed to implement the proposed model within the organization. Researcher has identified financial and human resources will be necessary to introduce some of the elements of the proposed Canvas, especially within the context of setting up and integrating an internal algorithmic audit system (outlined in the Evaluation and monitoring, internal audit section), as well as corporate training system. Resource for corporate education, namely introducing internal ethics and data management roles (i.e., ethics officer in CDO team or divisions dealing with knowledge management); workshops and lectures on the safe use and interaction with AS, e-learning system and related materials also serve as necessary requirements for the successful model implementation.

Value proposition (VP)

Value proposition (VP) is aligned with the Value proposition part of the Business Model Canvas (Alexander Osterwalder & Pigneur, 2010) and specifies the value of Algorithmic Accountability Canvas as a proposed model. Proposed artefact encompasses a set of tools, practices and guidance developed to improve algorithmic accountability within the organizational context. Algorithmic Accountability Canvas serves as a generalizable solution aimed at assisting organizations in designing accountable algorithmic systems in order to identify and prevent harmful outcomes from AS deployment and utilization. The value created from improving algorithmic accountability organization-wise is addressed in Value created section of the Algorithmic Accountability Canvas.

Stakeholder responsibility clarification (SRC)

Stakeholder responsibility clarification (SRC) is aligned with the Customer relationships of the original Business Model Canvas (Alexander Osterwalder & Pigneur, 2010) and represents stakeholder relationships and responsibility distribution as opposed to customer relationships due to internal organizational focus of the proposed model. Stakeholder responsibility clarification section addresses the necessity of distribution of associated responsibility for actors involved in designing, developing and deploying AS within the company. Organization should perform an explicit delegation of responsibilities and tasks between associates in charge of design of AS and algorithms. This section corresponds to the Design Principle of Stakeholder responsibility clarification (Table 8) and is aimed at achieving better awareness of associate's responsibility scope and responsibilities of their supervisors to avoid the culture of blame shifting as reflected in the EC8 and achieve clear vision of roles within the decision system.

Data (DT)

Data section of the Algorithmic Accountability Canvas was added to the BMC (Alexander Osterwalder & Pigneur, 2010) in order to address the specifics of algorithmic accountability concept as a problem domain. Data reflection remains one of the most relevant issues in the field of AI ethics and ethically aligned design of AS/AI (The IEEE Global Initiative on Ethics of

Autonomous and Intelligent Systems, 2019). Researcher proposes a set of tools based on the extant literature in the field of fairness, transparency and accountability relating to utilization of AS (Gebru et al., 2018; Mitchell et al., 2019). In order to facilitate dataset and model reflection, the following tools clarifying intended use cases and documentation of datasets and models are proposed: datasheets for datasets (Gebru et al., 2018) and models cards for model reporting (Mitchell et al., 2019). Datasheets were developed as an explanatory tool intended to accompany the dataset by clarifying its intended use case, motivation, collection process and composition to improve communication between the creators of the dataset and intended dataset consumers with the long-run objective of prioritization of accountability and transparency within the machine learning and AS communities (Gebru et al., 2018). Similarly to datasheets, model cards were proposed by a group of researchers in the field of fairness, transparency and accountability of AI/AS to clarify the intended use cases of machine learning models and minimize their usage in the contexts which may not be well suited for them (Mitchell et al., 2019). Model cards are brief documents, which are intended to accompany machine learning models in order to provide "benchmarked evaluation in a variety of conditions, such as across different cultural, demographic, or phenotypic groups (e.g., race, geographic location, sex, Fitzpatrick skin type and intersectional groups (e.g., age and race, or sex and Fitzpatrick skin type) that are relevant to the intended application domains" (Mitchell et al., 2019). Datasheets for datasets and model cards for model reporting together serve as explanatory and descriptive tools aimed at complementing existing transparency practices and strengthening algorithmic accountability organization-wise.

Transparency (traceability, explainability) (TSP)

Transparency section of Algorithmic accountability Canvas is aligned with the The principle of transparency (Table 7). It introduces a set of measures aimed at strengthening transparency in algorithmic practices within the organization. In the context of the current study, transparency also addresses the concepts of explainability and traceability in accordance with the Principles of ethically aligned design (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). Researcher emphasizes that opacity and inscrutability of AS do not exempt organizations from being accountable, as companies are accountable for the decisions that are difficult to explain (Martin, 2019). However, limitations for transparency within the AS domain linked to information disclosure and trade secrecy call for increased attention and careful consideration of appropriate practices. Moreover, some studies have previously accentuated lack of feasibility in achieving full transparency (Ananny & Crawford, 2018). Researcher proposes possible solution for the transparency dilemma outlined in Martin (2019), where transparency for algorithmic decision-making should be specifically targeted to the type of decision and purpose. Martin (2019) argues that "the transparency needed for corporate responsibility in the principal–agent relationship (a large role of the algorithm in a pivotal decision) would differ from the transparency needed for an algorithm that decides where to place an ad". Essentially, type and level of transparency chosen by an organization is a design decision by itself and implies associated responsibility and adherence to the norms of the decision context.

Evaluation and monitoring (internal audit) (IA)

The section of Evaluation and monitoring (internal audit) of the Algorithmic Accountability Canvas was added to the original BMC (Alexander Osterwalder & Pigneur, 2010) along with the Independent oversight (external audit) section in order to highlight the importance of auditing and compliance mechanisms in the accountability problem domain. Evaluation and monitoring section proposes an internal algorithmic auditing framework outlined in Raji et al. (2020) as a main tool to for internal algorithmic audit system implementation. The aforementioned study introduces an internal audit process structure including scoping, mapping, artifact collection, testing, reflection and post-audit phases in accordance with the internal framework for algorithmic auditing that supports AS/AI system development end-to-end (Raji et al., 2020). Internal algorithmic audit framework is intended to close the accountability gap in AI/AS development and deployment and to provide a comprehensive set of processes and activities to ensure audit system integrity (Raji et al., 2020). The framework serves as a method to hold organizations, which design, develop and deploy AI/AS accountable for system compliance and declared ethical norms and principles through rigorous algorithmic auditing process throughout internal organization development lifecycle.

Independent oversight (external audit) (EA)

Independent oversight (external audit) section of Algorithmic Accountability Canvas is a newly added component aligned with The principle of transparency (Table 8) and intended to highlight the specifics of algorithmic accountability problem domain and importance of audit mechanisms and monitoring in it. Even though Independent oversight cannot be considered as one of the internal organizational measures, it is necessary to emphasize the importance of independent oversight within the wider context of algorithmic accountability, especially considering the rapidly evolving legal landscape in the field of AI/AS utilization in organizations. It is important to note that independent oversight largely depends on the current legislation in the specific state or region and particular legislative norms the company has to adhere in that case.

Cost structure/budget (CS)

Cost structure section corresponds to Cost structure part of the original BMC (Alexander Osterwalder & Pigneur, 2010) and refers to potential costs resulting from realizing activities and establishing resources for the model proposed. Researcher has identified costs generally related corporate training and education activities (including costs of educating developers and designers of AS outlined in the Key activities section, as well as value-based design incentivization activities, e-learning system development, workshop and lectures on the safe use and interaction with AS).

Moreover, support tools development costs will be added to the current cost structure, including resources necessary for internal algorithmic audit system development, data clarification tools implementation, guideline and codes of conduct development. Finally, human resources costs will be added, including those incurred from establishment of new roles within the organization (i.e., ethics officers in CDO teams or divisions dealing with knowledge management).

Value created (VC)

Value created is a final component of Algorithmic Accountability Canvas, which corresponds with the Revenue streams part of the original BMC (Alexander Osterwalder & Pigneur, 2010). As the current tool is aimed at solving an internal organizational issue, researcher proposed Value created part to be better aligned with the specifics of the intended use of the current model. Researcher has identified the value created from improving algorithmic accountability divided into quantitative and qualitative parts. Quantitative value relates to financial risk management, such as mitigation of risks from potential cases of algorithmic bias and risks related to non-compliance to algorithmic audit instances and other compliance mechanisms. Qualitative value created can subsequently be divided into Brand image and recognition and Social value, where the former addresses securing consumer trust and serving as a facilitator for organization's competitive and brand strategy and the latter deals with contribution to collective human well-being and society through the means of following the principles and norms of ethically aligned design.

The Design Principles and Algorithmic Accountability Canvas discussed above serve as the key resulting artefacts for the Design stage of the ADR process. Our study will proceed with the Implementation phase in order to support the instantiation of aforementioned artefacts through implementation and evaluation activities in the case company.

4.3 Implementation

Implementation stage follows the Design part of the ADR project. Implementation stage is realized through instantiation of artefacts developed in the previous phases of the study through implementation activities within the client organization. This part of the study assists instantiation through ongoing intervention activities in order to evaluate efficiency of proposed artefacts. In accordance with the ADR methods, the resulting artefact in Implementation phase may include system instantiation or a process (Mullarkey & Hevner, 2019). Similar to the previous cycles of the ADR project, Implementation phase by itself includes Problem Formulation or Planning, Artifact Creation, Evaluation, Reflection and Learning stages. Implementation phase roughly corresponds to the second half of Building and Intervention (BIE) phase in the original ADR model (Sein et al., 2011).

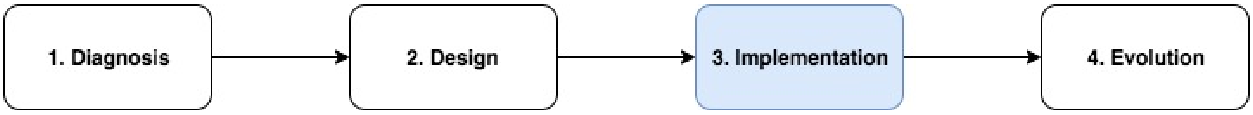


Figure 8. Elaborated Action Design Research process model cycles adapted from Mullarkey and Hevner (2019): Implementation stage

Following the thorough investigation of the problem domain and its importance during the Diagnosis phase, as well as development of artefacts during the Design stage, including Design Principles and Algorithmic Accountability Canvas, the current Implementation cycle serves as an evaluation engagement performed in the client organization. Extant literature presents various types of activities and approaches taken during the Implementation cycle within empirical studies relating to ADR methods. Previously, researchers have performed testing the concepts at a tradeshow in order to collect the feedback from the intended user groups (Schouten et al., 2020), implemented business intelligence and analytics cost allocation solution in a medium-sized company (Grytz et al., 2020), delivered a training program targeted at government officials addressing the problem of limited adoption of e-government services in developing countries (Gregor et al., 2014) and evaluated a prototype of an incident management system to support sensemaking and usability (van Wyk et al., 2020).

In the scope of the current ADR project, Implementation phase addresses the problem of improving algorithmic accountability as a part of the organizational IT strategy for the case firm and aims to evaluate the effectiveness and efficiency of Algorithmic Accountability Canvas as a proposed solution artefact. In order to achieve this goal, researcher has performed a series of practitioner workshops with the business strategy associates (Workshop A) and engineering associates (Workshop B) and facilitated follow-up discussions. This approach allowed for gathering evaluation evidence, including structured support and impressions from the practitioner side as well as identification of possible limitations and drawbacks of the proposed Algorithmic Accountability Canvas. The discussion below will address activities performed during the workshops.

Workshop A

The first participatory workshop was conducted within the business strategy, sales and marketing side of the ADR project team members. Initially workshop was planned to be held face-to-face in the company office, however, due to limitations inflicted by State of Emergency in Tokyo due to COVID-19 situation, it was decided to conduct the workshop online. Participating members were informed about the workshop a week prior. Additionally, researcher recorded the workshop to ensure continued access necessary for further data analysis. Participants assessed alpha prototype version of the Algorithmic Accountability Canvas and were encouraged to share their opinions, present ideas and openly discuss practical value of the proposed tool. Collected data was transcribed and open coded using qualitative data analysis tool NVivo similar to the procedure applied during Diagnosis part of the project. The resulting open coding framework is presented below.

Node	**Description**
Stakeholder resistance	Participants referred to examples from automotive industry, such as an Adaptive Cruise Control technology (ACC) to highlight the importance of stakeholder resistance to acceptance of various initiatives, including either new IS implementation or organizational changes (codes of ethics, standards, norms). In regard to educating designers and developers of A/IS and value-based incentivization of changing the narrative of neutral algorithms, similar problems may arise. Participants argue that in order to facilitate acceptance, top-down approach is necessary, such as government leading the key initiative.
Leveling transparency of A/IS depending on algorithmic decision scope and role in society	Participants unanimously agree on the significance and effectiveness of the proposed transparency/effort dilemma solution within Algorithmic Accountability Canvas. Participants argue that linking the role of A/IS in a decision to the role of algorithm's decision in society is an efficient way to ensure that no extraordinary additional effort is spent on conforming to transparency standards.
Cultural norms and ethical sensitivity	Participants point out that cultural norms within the specific region may affect speed of adoption and overall characteristics of legislation relating to algorithmic accountability. Participants referred to particular examples of specifics of cultural norms being different across regions, such as different levels of ethical

	sensitivity towards specific issues (e.g., racial intolerance) due to variety of socio-cultural and anthropological factors.

Table 10. Workshop A Nodes and corresponding description

Overall participant impression of Algorithmic Accountability Canvas was positive. Associates acknowledged the relevance of value proposition of the tool as a solution aimed at assisting case company in designing accountable algorithmic systems as a part of the organizational IT strategy. Particularly, respondents emphasized the importance of value-based design incentivization and education activities aimed at developers and designers of A/IS. Key activities (KA) were found to be an efficient way to realize the Value Created (VC) part of the canvas and additional potential ways to facilitate stakeholder responsibility clarification and develop support tools were also brainstormed during the workshop. Moreover, participants unanimously acknowledged the proposed approach to tackle the transparency/effort dilemma as a part of the canvas, highlighting the importance of context-based approach for transparency and role of algorithmic decision in society.

However, some drawbacks have also been identified. Participants have pointed out the issue of cultural norms and ethical sensitivity depending on specific region, especially in relation to their influence on value-based design methods to be used for algorithm development and deployment. Particularly, as the case company is located in Japan, participants would like to better understand how applicability of Algorithmic Accountability Canvas would depend on the speed of legislation adoption dealing with ethical and fairness issues for algorithms. Due to various socio-cultural and anthropological factors, some cultures tend to have different levels of ethical sensitivity towards specific issues (Chan & Cheung, 2012; Chung et al., 2008; Simga-Mugan et al., 2005). Speed of adoption for algorithmic accountability legislative norms may be slower in Japan due to specifics of cultural norms and other related factors. Moreover, participants argue that in countries like Japan, enforcement of legislation and top-down approach in facilitating new initiatives is especially important due to initial stakeholder resistance. As one of the respondents has mentioned:

"People need to feel acceptance towards this idea and the initiative should be led by the government. In general, in Japan businesses need a successful example of an idea introduced elsewhere first. For example, in automotive industry we have a technology named Adaptive Cruise Control technology (ACC). Japanese OEMs were hesitant to implement it, even though the solution was ready technology-wise. But as soon as OEMs in the United States introduced it, Japanese OEMs decided to follow right away."

Participants emphasize significance of standardization and its influence on stakeholder readiness to follow and accept new initiatives. Additionally identified limitations and participant comments are further considered as a basis for revisions of the proposed artefacts. In conclusion, Algorithmic Accountability Canvas is experienced as positive enabler for improving algorithmic accountability as a part of organizational overall IT strategy.

Workshop B

Workshop B was conducted as a part of the Implementation cycle with the engineer side of the ADR team. Workshop structure remained the same as Workshop A, allowing to gather evaluation evidence based on structured support and impressions from the practitioner side and identification of possible drawbacks of the proposed tool. Participants expressed their opinions regarding functionality of the proposed canvas and engaged in discussion about its configuration. Workshop B was conducted online due to limitations imposed by COVID-19 situation in Tokyo at the time. Researcher has received permission to record the workshop in order to have continued access to the data. Subsequently, the data was transcribed and analyzed in NVivo qualitative data analysis tool. The resulting open coding framework is outlined below.

Node	**Description**
Corporate education seen as an efficient method of improving algorithmic accountability	Respondents unanimously agreed that providing corporate training and developing support tools (Key Activities part of the Algorithmic Accountability Canvas) in an organization serves as an efficient method to educate developers, designers of A/IS and other stakeholders on societal impacts of A/IS utilization. Respondents shared their experience of participating in activities, similar to those outlined as a part of Key Activities and Key Resources segments of the canvas. Respondents believe that workshops and lectures on the safe use and interaction with A/IS will serve as a major facilitator of recognition for societal implications of A/IS as a problem space and improving algorithmic accountability overall.
Leveling transparency of A/IS depending on the algorithmic decision scope and role in society	Workshop B participant views regarding transparency/effort dilemma and corresponding proposed solution outlined in the Algorithmic Accountability Canvas (Transparency, Traceability and Explainability part) aligned with the views of respondents from Workshop A. Respondents agree that linking the role of an algorithm in a decision (ranging from small to large) to the role of algorithm’s decision in the society (from minimal to pivotal) is an effective potential solution to tackle the issue of ensuring

	transparency without requiring excessive effort on the part of a company, potentially putting a strain on organizational resources.
Internal and external audit interrelation	Participants pointed outed that internal audit should be based on the assumption that external audit exists, "so that internal audit can follow along." Similar to the views reflected in the Workshop A, participants in the Workshop B stress the importance of legislation and its maturity in regard to the problem domain. While participants agree that internal audit is necessary for ensuring compliance with the internal codes of ethics and other related internal normative base, government level initiative is seen as a key factor for facilitating problem domain consciousness.

Table 11. Workshop B Nodes and corresponding description

Workshop B reflected overall positive participant impression of Algorithmic Accountability Canvas as a tool aimed to improve algorithmic accountability within an organization. Similar to participant views captured during Workshop A, respondents emphasized the relevance of the idea of leveling transparency of A/IS depending on the algorithmic decision scope and role in society (TA part of the Algorithmic Accountability Canvas). Moreover, researcher has identified a shift in opinions regarding necessity of addressing a wide range of stakeholders to improve understanding of ethical and societal implications of using algorithmic systems. Since some participants of Workshop B were the engineers interviewed during the Diagnosis stage of the ADR project, researcher could track the difference between the initial perception of ethical implications of A/IS as a problem domain and related issues (AI ethics in general, educating stakeholders about how fairness can be introduced into A/IS design) and views regarding relevance of educating stakeholders about societal impacts of A/IS during the Implementation phase. In accordance with the data collected during the Diagnosis stage, engineers expressed their skepticism regarding the efficiency of corporate training to improve algorithmic accountability, e.g.:

"Some people just develop the interface, just the looks of it, they do not work specifically with the algorithms. Not every single stakeholder has to take those classes…"

However, Workshop B respondents agreed that providing corporate training and developing support tools (Key Activities part of the Algorithmic Accountability Canvas) in an organization serves as an efficient method to educate developers, designers of A/IS and also other stakeholders on societal impacts of A/IS utilization and subsequently serves as one of the measures to improve algorithmic accountability. Within the scope of the ADR project researcher continuously facilitated discussions with the ADR team members about the problem domain and provided real business and societal examples of A/IS unfavorable impacts. Implementation phase of the ADR project reflected that the views of the engineers have shifted towards a clearer understanding of importance of ethical implications and indirect biases in the AS design problem space. Based on the Implementation phase Workshop B data, we can conclude that Key Activities (KA) part of the Algorithmic Accountability Canvas was proven to be an efficient enabler of improving algorithmic accountability in an organization.

Workshop A and Workshop B frameworks serve as a concluding artefact for the Implementation stage of ADR research project. The study will continue with the Evolution stage in order to differentiate the initial design context from the successive evolutionary design context.

4.4 Evolution

Evolution serves as a concluding stage of the elaborated ADR project. Evolution stage is realized through the process of addressing previously instantiated artefact and specifically its evolution over time. The last e-ADR project cycle addresses potential problem environment changes and how the proposed artefact solution evolves to tackle these changes. According to e-ADR methods (Mullarkey & Hevner, 2019), evolutionary processes and interventions such as design improvements and technological developments may be a long-term organizational initiative and will continue to contribute to knowledge generation useful to both the researcher and practitioners alike.

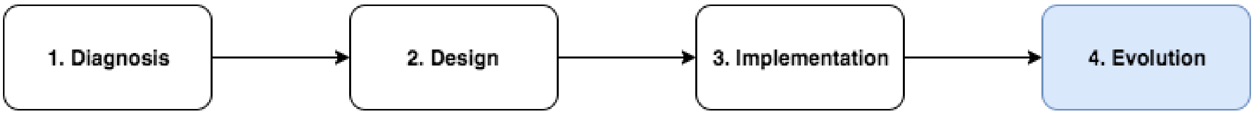


Figure 9. Elaborated Action Design Research process model cycles adapted from Mullarkey and Hevner (2019): Evolution stage

Evolution stage addresses the issue of reconsidering the previously instantiated artefacts sometime after the Implementation stage is concluded in order to assess how they may perform over time. Typical resulting artefact during the Evolution stage may include an improvement or a new artefact of any of the artefact types within the three first stages of the ADR project. Moreover, e-ADR method posits that project separation into smaller chunks allows for better project management and well-defined sequence of activities in order to ensure better control of the overall project planning and development. Therefore, Evolution stage provides an opportunity for differentiation between the initial design context from the subsequent design context. In accordance with the original ADR method description outlined in Sein et al. (2011), the final stage of the ADR project Formalization of Learning addresses the process of formalization of the design principles, articulation of the class of problems and class of solutions and determining the ensemble specific knowledge.

Within the scope of the current ADR project, Evolution stage is realized through assessing how the proposed artefact may evolve with time. Researcher revisits the instantiated artefacts and addresses the limitations and drawbacks identified throughout the project interventions. The refined artefacts will serve as an addition to the knowledge base for both the theoretical and practical streams. Evolution stage also ensures the process for reporting knowledge contribution is comprehensive and is aligned with the DSR and ADR Design Principles, including the principle of Research Rigor, which posits that research relies upon the application of rigorous methods in both the development and evaluation of the design artefact (Hevner et al., 2004).

As the current study utilized a problem-centered entry point strategy, researcher started with the initial problem identification during the Diagnosis stage. After conducting in-depth investigation of the problem domain during the Diagnosis phase, researcher performed building and evaluation activity for a set of design principles to create a tool that would serve an enabler for improving algorithmic accountability within the organization. Upon completion of the Design stage, instantiation of the newly developed artefacts was conducted in the case company, allowing

ADR team to evaluate the proposed artefacts through an intervention with the firm and produce learning and reflections. Evolution is the last stage of the ADR project in accordance with the ADR methods and focuses on assessing how the proposed artefact evolves over time as the problem environment changes. However, researcher has identified some shortcomings related to e-ADR process model description of the concluding project cycle. Specifically, Mullarkey&Hevner (2019) argue that the essence of the Evolution stage is «a need to re-consider instantiated artefacts at some point after implementation and during or after adoption as to how they evolve over time» (p.10). Nevertheless, no further guidance is provided in relation to specific activities to be performed within the Evolution stage in order to achieve this objective. In comparison to more comprehensive approach in describing and providing support in realization for the preceding e-ADR phases (Diagnosis, Design and Implementation), we find that e-ADR model description for the Evolution cycle lacks thorough guidance and provides excessive variability in interpretation of methods and activities to utilize while conducting e-ADR. However, it contradicts the initial aim of e-ADR to «fully elaborate and actualize the ADR process model in order to aid the conduct of each intervention cycle…better support users to structure the key decisions and activities necessary to rigorously apply ADR» (Mullarkey & Hevner, 2019, p.6). Moreover, we argue that Evolution stage is disconnected from Formalization of learning activity outlined in the original ADR process model, which prescribes articulation of class of problems, class of solutions in order to satisfy ADR generalization principle (Sein et al., 2011). Therefore, we propose a schematic representation of activities to be performed during Evolution phase in order provide a more comprehensive, elaborated approach to enhance the experience of e-ADR methods application.

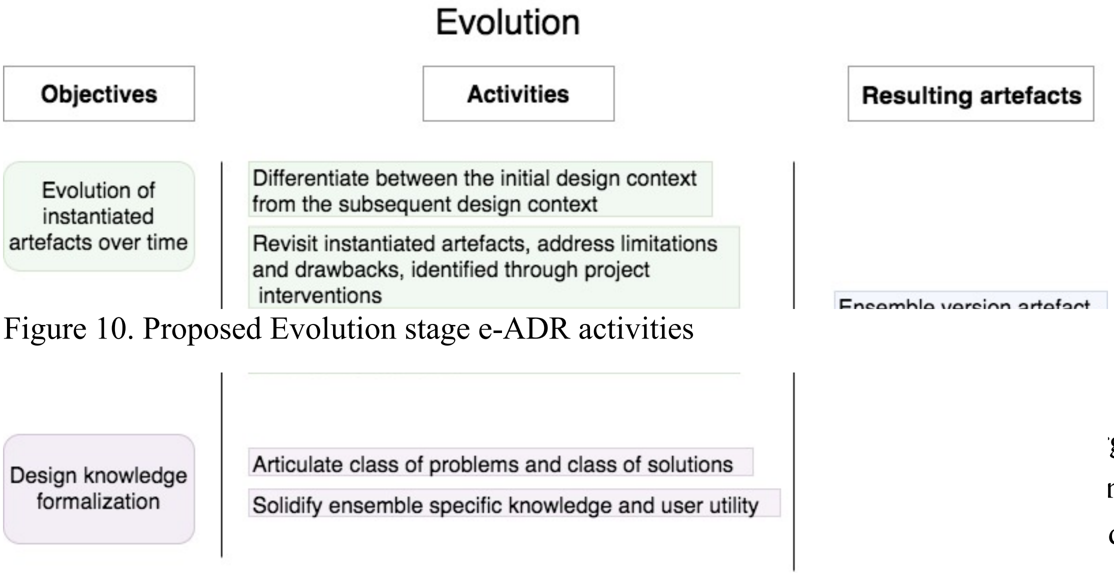

Evolution
Objectives
Activities
Resulting artefacts
Evolution of instantiated artefacts over time
Differentiate between the initial design context from the subsequent design context
Revisit instantiated artefacts, address limitations and drawbacks, identified through project interventions
Ensemble version artefact
Design knowledge formalization
Articulate class of problems and class of solutions
Solidify ensemble specific knowledge and user utility

Figure 10. Proposed Evolution stage e-ADR activities

ganization,
nt changes.
d. Artefact

instantiation has revealed that cultural norms within the specific region may affect speed of adoption and overall characteristics of legislation relating to algorithmic accountability. Particularly, speed of adoption for algorithmic accountability legislative norms may be slower in Japan due to specifics of cultural norms and other related factors. Moreover, in countries like Japan, enforcement of legislation and top-down approach in facilitating new initiatives is especially important due to initial stakeholder resistance. These issues are addressed through revision of the original Design Principles by adding the principle of cultural awareness and ethical sensitivity, which represents appreciation of the potential environment change and difference between the initial and subsequent design contexts. The principle of cultural awareness and ethical sensitivity is aimed to provide necessary revision derived from unanticipated consequences based on the feedback throughout the implementation activity in case organization.

b) Design knowledge formalization

Contributing to the existing scientific body of knowledge on improving algorithmic accountability within an organization as a part of its IT strategy, we have acquired design knowledge within the scope of an elaborated ADR project formalized in a set of design principles. The design principles embody our findings throughout an extensive practice-based project and our generalized knowledge for the solution we have developed within the project duration. We formalized the design knowledge by the means of continuous process of artefact revisions and their evaluation, participative workshops, application of analytical memo writing and analysis of the extant literature. We claim that the derived revised version of design principles for improving algorithmic accountability within an organization satisfies the generalization principle of ADR by articulating the class of problems (algorithmic accountability), class of solutions (business model canvas tool for improving algorithmic accountability in an organization) and associated design principles. The finalized design principles provide dual utility by contributing to the existing scientific knowledge base in design research field, as well as embodying emerging design knowledge for practitioners utilizing algorithmic systems. We also claim that scientific research contribution of this study is justified by providing the use of e-ADR method within the specific class of problems context and contributing to the IS-literature by highlighting the socio-technical nature of the algorithmic accountability phenomenon. Some of the limitations of the current study are related to deriving results based on the project conducted within the context of the single case company. This point will be further addressed in the discussion part of the study.

Design principle	**Explanations and rationale**

DP1 The principle of raising awareness of ethics and ethical literacy	Limited understanding of ethical implications of using algorithms and misalignment between the state of research in the field of ethically aligned AI systems (including industry-produced Codes of ethics) and the practical state of participant engineers' awareness of ethics. It is necessary to educate stakeholders on societal impacts of designing, developing and implementing AS.
DP2 The principle of value-based design incentivization and appreciation of AS deployment context	Values-based design methods should be put in the center of the technical system development in order to create sustainable systems providing not only economic value to the organizations but increasing human and societal well-being. In order to ensure that the algorithms in organizations are utilized responsibly, stakeholders need to consider the notion of value-laden algorithms, as opposed to free of bias, neutral narrative of algorithm usage.
DP3 The principle of actionable guidelines	Industry guidelines for A/IS ethics should include actionable statements rather than descriptive principle and value-related formulations, which are too vague to translate into tangible results. Incorporating ethics into technology design agenda also deals with translating the norms into language accessible to different levels of stakeholders (e.g., policy nuances into technical context).
DP4 The principle of transparency	Lack of transparency increases the difficulty in achieving accountability (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). Operation of AS should be made transparent to a wide range of stakeholders, however transparency (also addresses explainability and traceability) may need to be targeted towards specific type of decision and purpose (Ananny & Crawford, 2018).
DP5 The principle of stakeholder responsibility clarification	Design of algorithms calls for clear understanding of responsibilities and roles of the decision system (Martin, 2019). Clarifying participant dynamics helps to ensure more transparent provision of information and improved interpretation of the system usage context.
DP6 The principle of cultural awareness and ethical sensitivity	Cultural background of an individual is known to have a significant effect on her ethical sensitivity (Blodgett et al., 2001; Fernando & Chowdhury,

	2010). Various cultures (e.g. collectivist of individualistic) impact people's perception of ethical dilemmas and behavior of individuals in organizations (Husted & Allen, 2006). Since ethical sensitivity of an individual (associate) may differ depending on socio-cultural norms in specific region (company location), it is necessary to consider related factors, such as speed of algorithmic accountability legislation adoption and varying levels of stakeholders' ethical sensitivity to specific issues.

Table 12. Revised design principles and corresponding explanations and rationale

Revised design principles serve as a resulting artefact for the Evolution stage of the e-ADR project. Evolution stage marks the conclusion of the empirical part of our research. Our study will proceed with the discussion of results, limitations and implications for future research.

Chapter 5. Discussion

5.1 Theoretical contribution

Our study aimed to contribute to theoretical stream of knowledge by obtaining design knowledge within the scope of an elaborated ADR project formalized in a set of design principles for improving algorithmic accountability within an organization. On the theoretical side, this study aimed to add to the IS literature base by reflecting the sociotechnical nature of algorithmic accountability phenomenon and building ensemble design knowledge for organizations utilizing algorithmic systems. We achieved the aforementioned goals by applying e-ADR process model (Mullarkey & Hevner, 2019) in order to iterate nascent design theory to inform artefact design and use across problem domain in question (algorithmic accountability). Moreover, we attempted not only to inform research and practice by developing an innovative artefact for specific contextual use, but also to demonstrate its utility across the whole class of field problems domain.

From a theoretical point of view, our study is concerned with a class of problems that, to our knowledge, has not been empirically addressed before as a problem domain in action design research field, despite identified fitness of ADR methods for developing socio-technical design agenda for a specific class of problems (Sein et al., 2011). We managed to close the identified gap by applying e-ADR method within the scope of an extensive case study of a large MNC located in Japan. We have achieved the initial goal of building prescriptive design knowledge and contributing to IS theory formalized in a set of design principles for an instrument aimed to improve algorithmic accountability in an organizational context an proved its efficiency through instantiation activity in the case firm.

Building upon extant research on algorithmic accountability (Buhmann et al., 2020; Clavell et al., 2020; Diakopoulos, 2015) and ethically aligned design theory (The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019; Vakkuri & Abrahamsson, 2018; Weng & Hirata, 2018), identified as the kernel theories in this study, we derived design principles applicable to firms utilizing A/IS systems. We conducted in-depth investigation of the problem domain during the Diagnosis phase, collaborated with practitioners during an iterative process of developing a tool serving as an enabler for improving algorithmic accountability and performed instantiation of the newly developed artefact in the case company. Finally, we evaluated the proposed artefact through an intervention with the firm and produced learning and reflections. We contribute to the IS knowledge base by extending the nascent design theory (Markus et al., 2002) and articulating class of problems and class of solutions within the emerging field of algorithmic accountability (Wieringa, 2020).

Moreover, we have identified some shortcomings of elaborated action design research process model (Mullarkey & Hevner, 2019) description of activities within the concluding project cycle. To solve the identified issue, we proposed a schematic representation of activities to be performed during Evolution phase in order provide a more comprehensive, elaborated approach to enhance the experience of e-ADR methods application.

Prior research suggests that an effectively formulated design principle should contain the following three components: «First, information about the actions made possible through the use of an artifact. Second, information about the material properties making that action possible. Third, the boundary conditions under which the design will work» (Chandra et al., 2015, p. 4044). Therefore, design principles should provide prescriptive design knowledge about action, boundary conditions and material properties of the artefact. We claim that all the three conditions have been satisfied in our study, as the formulated design principles fall under action and materiality-oriented category in accordance with primary orientations provided in Chandra (2015), describing both how the system should be designed and what actions should it allow for. Moreover, our study specified the scope (boundary conditions) of the artefact, emphasizing on socio-technical dimension of algorithmic accountability as a phenomenon of socio-technical nature, embodying both technical and socio-cultural aspects to it. We specified that in the scope of the current study, algorithmic system is not recognized as a purely technical construct due to the fact that individuals involved in the design and development of the algorithmic system cannot simply be separated from its decisions and bias can find its way into the algorithms due to the many ways these individuals stay involved in the algorithmic decisions (Martin, 2019).

Finally, the issue of generalizability is one of the relevant topics within the field of IS and ADR in particular (Gregor, 2006; Lee & Baskerville, 2003; Purao et al., 2002) and therefore is addressed in the study. Essentially, due to highly situated nature of ADR generalization of outcomes may be challenging. Principle 7 «Generalized Outcomes» of the ADR method posits that the shift from the contextual to abstract and generic is one of the key components of ADR (Sein et al., 2011). Three levels are suggested in order to perform the move: generalization of solution instance, generalization of problem instance and derivation of design principles from the outcomes. However, ADR method does not specify or explicitly prescribe the number of case studies to be performed in order to attain generalizability. Various prior ADR studies have demonstrated achieving generalizable research outcomes based on the data collected in a single case company (Dremel et al., 2020; Reibenspiess et al., 2020). We claim that our study satisfies the requirements of ADR methods based on the systematic, theoretically grounded approach followed and having articulated class of problems, class of solutions and associated design principles. However, it is not feasible to assure that the results are exhaustive due to contextual, specific setting of current research. Further research may build up on design knowledge developed in this study by evaluating and extending presented insights in different contexts in order to iterate and inform the nascent design theory in the problem space of algorithmic accountability.

The table below represents summary of contributions of the study. We addressed design principles, algorithmic accountability concept, e-ADR method articulation and artefact contribution and provided corresponding description in regard to descriptive or prescriptive type of knowledge created.

Contribution	Description
Design principles	*Prescriptive knowledge* Design principles (regarding action, materiality and boundary conditions) for improving algorithmic accountability within an organization
Algorithmic accountability	*Descriptive knowledge* Provided articulation of algorithmic accountability as a class of problems, unpacked socio-technical view and highlighted socially constructed aspects of algorithmic accountability as a phenomenon as opposed to purely technical view of the concept
e-ADR method	*Prescriptive knowledge* Extended e-ADR method by providing representation of activities to be performed during Evolution phase
Artefact contribution	*Prescriptive knowledge* Developed, performed instantiation activity, evaluated and provided reflection on Algorithmic Accountability Canvas and derived associated design principles. Established proposed artefacts as a class of solutions type aimed to solve an identified organizational problem.

Table 13. Research contributions summary

Theoretical contributions part of the study concludes articulation of knowledge created within theoretical stream. The following part of the study will discuss practice-based knowledge contribution and managerial implications.

5.2 Managerial implications

This study was conducted in a real business setting within the context of a case company, Japanese branch of the German automotive and technology MNC. Due to iterative, collaborative nature of ADR method, we formed an ADR team consisting of researcher and practitioners in order to actively involve a number of relevant stakeholders, including engineering, R&D and business side associates. This study aimed to contribute to practice-based knowledge through an elaborated ADR method by addressing a real business problem and producing a set of design principles for the management in order to assist case organization in designing accountable algorithmic systems and improving currently realized algorithmic accountability practices. We also aimed to provide a generalizable solution for improving algorithmic accountability for businesses deploying A/IS, suitable for use outside of situational context outlined in the current study.

From a practical point of view, our study contributes by proposing design guidance for organizations utilizing A/IS allowing to improve algorithmic accountability as a part of the organizational IT strategy. As a rapidly growing number of companies are actively involved in designing and deploying algorithmic systems and more A/IS have an increasing impact on people's lives on the daily, the firms come under attention following the growing concern over accountability of such systems. Proposed and evaluated artefact encompasses a set of tools, practices and guidance developed to improve algorithmic accountability within the organizational context. Algorithmic Accountability Canvas serves as a generalizable solution aimed at assisting organizations in designing accountable algorithmic systems in order to identify and prevent harmful outcomes from AS deployment and utilization. The results of our study confirmed efficiency of the Algorithmic Accountability Canvas as an enabling tool for improving algorithmic accountability. Algorithmic Accountability Canvas serves as a blueprint for practitioners (e.g., managers) which can be optimized for the use depending on the specific organizational context (e.g. resources for the internal audit system development, cost structure and budget, socio-cultural and location factors). Additionally, formulated design principles embody managerial implications and insights derived from their evaluation and revision processes.

First of all, our study revealed that the engineers tend to be skeptical towards ethical implications of using A/IS. Particularly, our research reflected that even though bias and unfairness are recognized as a problem space in the algorithmic domain, it is perceived to be outside of the practical scope, relatively less important than maximizing the efficiency of an algorithmic system and challenging to implement. We formulated the principle of raising awareness of ethics and ethical literacy (DP1) and associated Key Activities (KA) within the Algorithmic Accountability Canvas to assist organizations in educating developers and designers of A/IS on the societal impacts of algorithmic systems, as well as improving an understanding of how fairness can be introduced into A/IS design. Secondly, our study shows that engineers tend to perceive algorithmic systems at a face value level of objectivity, forming a narrative of algorithms free of bias by default, which have to be «deliberately sabotaged» to produce unanticipated, unwanted outcomes. Nevertheless, prior research indicates that firms should be mindful of indirect biases, because ethical consequences of using algorithms are not necessarily pre-fixed in the design of the algorithmic systems (Martin, 2019). We formulated the principle of value-based design incentivization and appreciation of AS deployment context (DP2) and associated Value-based incentivization part within Key Activities (KA) of the Algorithmic Accountability Canvas in order to help companies integrate the concept of value-laden AS as opposed to "neutral" narrative of algorithms to mitigate the associated risks and assist them in putting value-based design methods in the center of the technical system development. Moreover, we developed the principle of actionable guidelines (DP3), which stipulates that industry guidelines for AI/AS ethics should include actionable statements rather than descriptive principle and value-related formulations, which are too vague to translate into tangible results. Furthermore, we developed the principle of transparency (DP4), as review of extant literature on the problem domain revealed that lack of transparency increases the difficulty in achieving accountability (Ananny & Crawford, 2018; The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019; H. J. Watson & Nations, 2019). We proposed a possible solution for the transparency and corresponding effort from the organizational side necessary to achieve transparency, which is based on linking the role of A/IS in a decision to the role of algorithm's decision in society as an efficient way to ensure that no extraordinary additional effort is spent on conforming to transparency standards. Additionally, we addressed the identified issue of lack of awareness regarding how responsibility is distributed between the stakeholders within the case company and formulated the principle of stakeholder responsibility clarification (DP5) and corresponding SRC part of the Algorithmic Accountability Canvas, which assists organizations in establishing an explicit delegation of responsibilities and tasks between associates in charge of design of AS and algorithms. Finally, we identified the need to revise the alpha version artefact during Implementation phase of the study and added the principle of cultural awareness and ethical sensitivity (DP6), which calls for consideration of cultural dimensions and diversity of the existing cultural norms in specific organization due to difference in ethical sensitivity of its stakeholders.

In conclusion, based on our evaluation results, Algorithmic Accountability Canvas serves as an enabling tool for organizations utilizing A/IS aiming to improve algorithmic accountability in a form of practical, easily understandable and directly translated into the business language set of tools and practices.

5.3 Limitations and future research

In the following part we wish to reflect on potential limitations of this study and provide opportunities for follow-up research. Due to immersive nature and researcher-practitioner collaboration of an ADR project, research activities were realized in iterative manner and engaged a number of representative stakeholders in order to build and evaluate an artefact aimed to solve an identified organizational problem of improving algorithmic accountability. We reached our initial research objective by proposing design guidance for organizations utilizing A/IS allowing to improve algorithmic accountability as a part of the organizational IT strategy. However, there are several limitations, which need to be addressed in the light of interpretation of our research results.

Firstly, our study was conducted within a single case company context, which may have affected generalizability of research outcomes. We have discussed the issue if generalizability and its relevance in ADR methods in detail within 5.1 Theoretical contribution part of this study. Since ADR method posits that the artefact emerges from interaction with the organizational context during its development and use (Sein et al., 2011), we would like to acknowledge the fact that we developed and evaluated the artefact based on the involvement with one single case company and investigating other contexts and types of companies could potentially lead to varying outcomes. Moreover, engagement within a specific company context may be limited to specific attitudes and experiences of associates of the case company. However, we would like to point out that ADR method does not specify or explicitly prescribe the number of case studies to be performed and prior ADR studies have demonstrated achieving ensemble artefact creation based on the data collected in a single case company (Dremel et al., 2020; Haj-Bolouri, 2019; Reibenspiess et al., 2020). Nevertheless, we are unable to claim that the results of our ADR project are exhaustive due to contextual setting of research. Despite the single case company limitation, our study contributes to novel design knowledge domain through applying ADR, since, to our knowledge, algorithmic accountability has not been empirically addressed before as a problem domain in ADR field in spite of identified fitness of ADR methods for socio-technical artefacts development.

Secondly, as the ADR project intent was guided by the researcher and the analysis was grounded in analytical memo writing, interview data collection and other qualitative collaborative methods, our research outcomes may suffer from bias related to subjectivity of involved parties. We also would like to invite follow-up quantitative research in order to gain comprehensive evaluation and confirm generalizability of our outcomes.

Our study lays a foundation for future ADR studies relating to the emerging field of algorithmic accountability. The future research may address the broader class of solutions and extend the proposed Algorithmic Accountability Canvas, as well as investigate larger contexts and involve wider groups of stakeholders. Moreover, future research may also concern a potential linkage between company brand image, reputation and efficiency in implementation of ethically aligned A/IS systems. Our study reflected that engineers tend to view responsibility from pragmatic and financially oriented viewpoint, linking organizational ethical considerations to purely compliance-related incentives. Participants also argued that maximizing revenue and reputational concerns serve as the primary objectives for the company to stay compliant, while ethical considerations will be an area of secondary importance. This discussion is out of scope for our study; however, future research may address the aforementioned insights.

Chapter 6. Conclusion

The objective of this dissertation was to address the problem of improving algorithmic accountability through developing a set of design principles in an organizational context. The overall research aim was to solve an organizational problem of improving algorithmic accountability by building an innovative artefact. To conclude the discussion on the study, we would like to revisit the original research questions and articulate how the results are enacted.

- RQ1: What are the appropriate design principles for improving algorithmic accountability in the organizational context?

In order to answer the main research question, we conducted an immersive practice-based study through applying an elaborated Action Design Research method as our research method of choice due to identified fitness of ADR for investigation and development of socio-technical artefacts. As algorithmic accountability is deemed to be a concept of a socio-technical nature, it calls for an interdisciplinary and collaborative approach in its investigation as opposed to a single-sided technocentric view (Wieringa 2020). We collaborated with practitioners and involved a number of stakeholders throughout our project, which unfolded in an organizational context of the Japanese branch of a globally operating technology company. We conducted in-depth investigation of the problem domain during the Diagnosis phase, collaborated with practitioners during an iterative process of developing a tool serving as an enabler for improving algorithmic accountability and performed instantiation of the newly developed artefact in the case company. Finally, we evaluated the proposed artefact, produced learning and reflections and obtained design knowledge formalized in a set of design principles. The revised set of design principles includes DP1 The principle of raising awareness of ethics and ethical literacy, DP2 The principle of value-based design incentivization and appreciation of AS deployment context, DP3 The principle of actionable guidelines, DP4 The principle of transparency, DP5 The principle of stakeholder responsibility clarification and DP6 The principle of cultural awareness and ethical sensitivity, discussed in detail within the Design, Implementation, Evolution and Discussion parts of the study.

To address the four additional research questions, which were developed in order to provide further support in addressing the main research question, we would like to revisit them and discuss how the study outcomes are enacted.

- How is algorithmic accountability realized in a case organization and what factors can serve as either facilitators or barriers for achieving it?

We investigated what factors can serve as barriers or facilitators for improving algorithmic accountability within the case organization during Diagnosis phase of the e-ADR project by performing a set of interviews with the company associates, as well as discussing the problem domain with the ADR team practitioner side representatives (business strategy side managers). We subsequently analyzed the data and developed an analytical framework in a form of Empirical Claims, which embodies obtained knowledge on main barriers for improving algorithmic accountability, among which is engineers' skepticism towards AI ethics, limited understanding of how fairness can be introduced through AS design, lack of awareness about AI ethics and ethics related internal guidelines and limited understanding regarding hierarchy in responsibility levels within the organization and personal accountability scope.

- What are the critical design principles and features for facilitating algorithmic accountability in a case company?

During Design phase of the project we have constructed alpha version artefact in a form of design principles for improving algorithmic accountability, which was further revised after instantiation of the alpha version artefact was performed in the case company. We articulate the development process for Design Principles within Design, Implementation, Evolution and Discussion parts of the study.

- How does the instantiated artefact (set of design principles) help to solve the identified problem?

During Implementation phase of the ADR project we performed an instantiation of the previously developed artefact (Algorithmic Accountability Canvas) through a series of case company workshops aimed to evaluate efficiency of the proposed solution. The results of our study confirmed efficiency of the Algorithmic Accountability Canvas as an enabling tool for improving algorithmic accountability. Implementation phase of the ADR project reflected that the views of the engineers have shifted towards a clearer understanding of importance of ethical implications and indirect biases in the AS design problem space, proving the effectiveness of the proposed artefact in realizing its declared value proposition.

- How can a problem solution generalization for achieving accountable algorithmic systems be developed?

The issue of problem solution generalization is discussed in detail within Theoretical contribution and Limitations part of the study. Principle 7 «Generalized Outcomes» of the ADR method posits that the shift from the contextual to abstract and generic is one of the key components of ADR (Sein et al., 2011). Three levels are suggested in order to perform the move: generalization of solution instance, generalization of problem instance and derivation of design principles from the outcomes. We claim that our study satisfies the requirements of ADR methods based on the systematic, theoretically grounded approach followed and having articulated class of problems, class of solutions and associated design principles.

Appendices

Figure 11. Action Design Research Process Model based on Mullarkey and Hevner (2019)

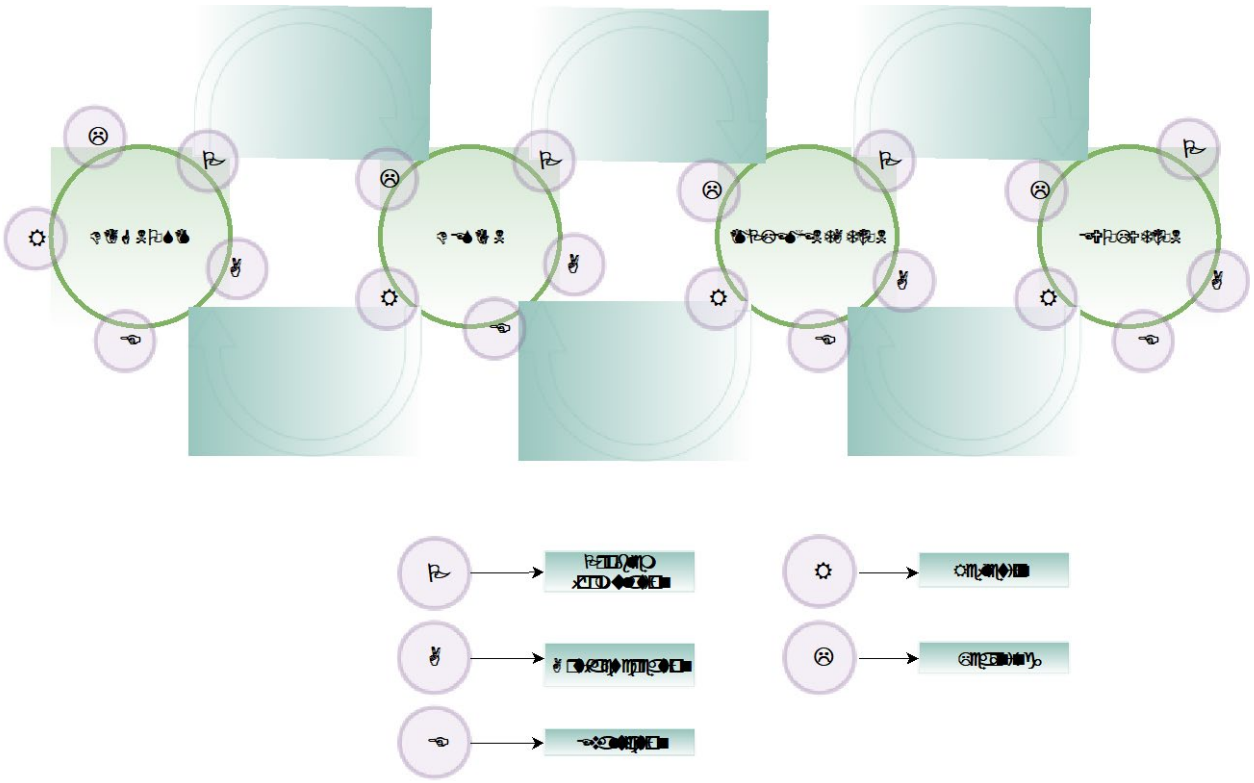
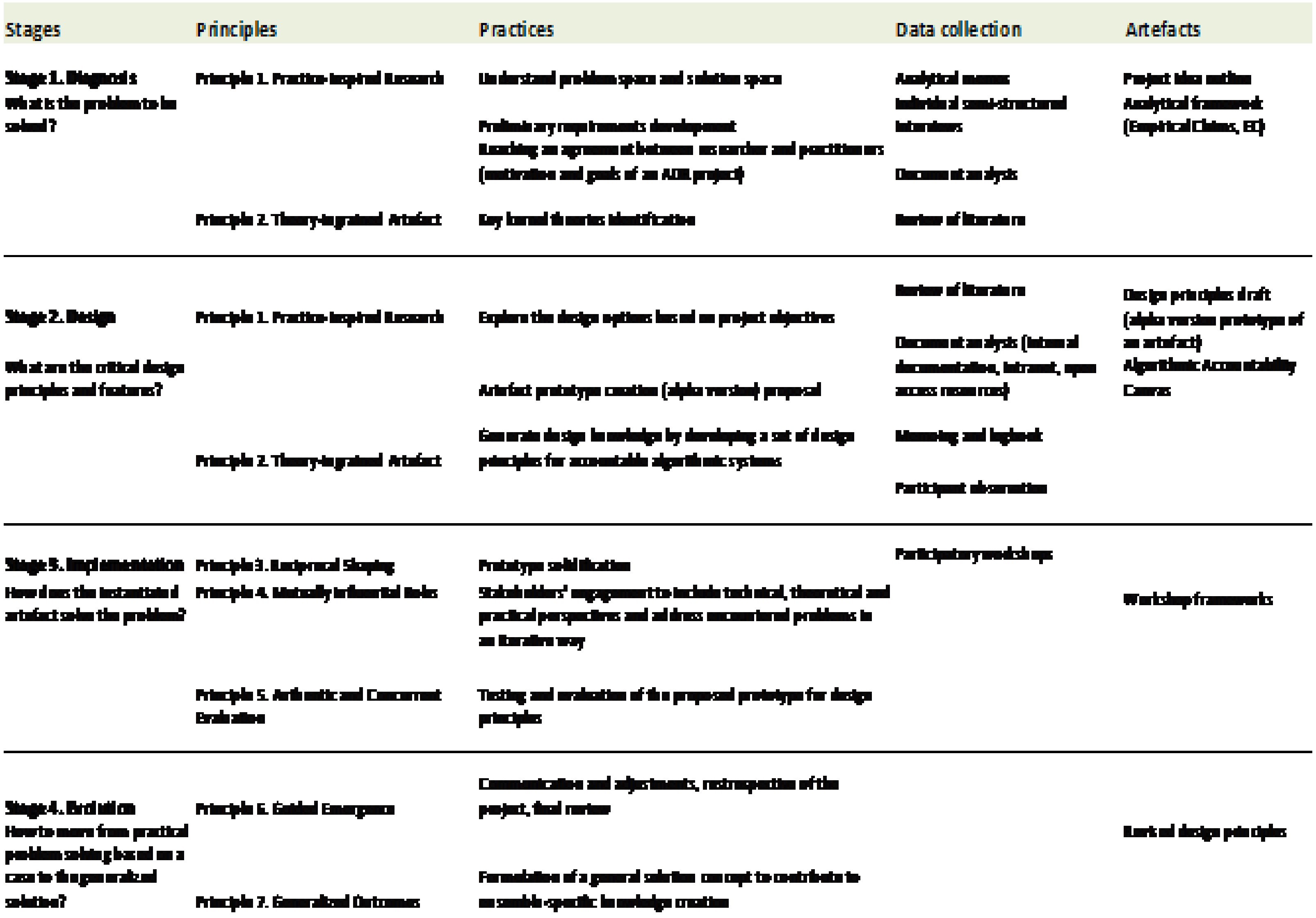


Figure 12. Work Breakdown Structure (WBS)



Stages	Principles	Practices	Data collection	Artefacts
Stage 1. Diagnosis What is the problem to be solved?	Principle 1. Practice-Inspired Research	Understand problem space and solution space Preliminary requirements development Reaching an agreement between researcher and practitioners (motivation and goals of an ADR project)	Analytical memos Individual semi-structured interviews Document analysis	Project idea outline Analytical framework (Empirical Claims, EC)
	Principle 2. Theory-Ingrained Artifact	Key kernel theories identification	Review of literature	
Stage 2. Design What are the critical design principles and features?	Principle 1. Practice-Inspired Research	Explore the design options based on project objectives Artifact prototype creation (alpha version) proposal	Review of literature Document analysis (internal documentation, intranet, open access resources)	Design principles draft (alpha version prototype of an artefact) Algorithmic Accountability Canvas
	Principle 2. Theory-Ingrained Artifact	Generate design knowledge by developing a set of design principles for accountable algorithmic systems	Memoing and logbook Participant observation	
Stage 3. Implementation How does the instantiated artefact solve the problem?	Principle 3. Reciprocal Shaping Principle 4. Mutually Influential Roles	Prototype solidification Stakeholders' engagement to include technical, theoretical and practical perspectives and address encountered problems in an iterative way	Participatory workshops	Workshop frameworks
	Principle 5. Authentic and Concurrent Evaluation	Testing and evaluation of the proposed prototype for design principles		
Stage 4. Evolution How to move from practical problem solving based on a case to the generalized solution?	Principle 6. Guided Emergence	Communication and adjustments, retrospection of the project, final review		Revised design principles
	Principle 7. Generalized Outcomes	Formulation of a general solution concept to contribute to ensemble-specific knowledge creation		

Appendix A. Interview disclaimer and questions

Interview disclaimer

As the mobility industry becomes significantly influenced by AVs, issues relating to implementation of accountability, fairness and transparency into autonomous systems becomes increasingly important.

Our research aims to explore the concept of algorithmic accountability, which generally can be described as the principle that an algorithmic system should employ a variety of controls to ensure the operator can verify it acts in accordance with its intentions, as well as identify and rectify harmful outcomes (New & Castro, 2018)

A gap between research and practice in the field of AI accountability and ethics can be currently observed. For example, a recent study by McNamara et al. (2018) revealed that Association for Computing Machinery (ACM) Code of Ethics had no observed effect on the way developers work. This shows that even though various guidelines for implementing ethics into software development exist, in practice they are not used within the industry.

Focus of the interview: to identify current practices, tools and methods through which algorithmic accountability is realized in the case company, as well as concerns (if any) of relevant stakeholders and accountability implementation in developing AS.

Interview questions

1. Do you think that it is necessary for developers working with AI to make the systems they develop “fair” by design?

2. How do you understand the following phrase: «algorithmic accountability»?

3. In a case when an algorithmic system utilized in an organization is found to produce faulty outcomes (e.g., bias), who do you think should be held accountable?

   Example: A third-party company is hired by a public agency to develop an algorithm for justice system. However, after it was developed and widely applied, an independent investigative company ran an analysis and found out that the algorithm made systematic racially discriminating mistakes (it is biased). Who should be held accountable?

4. During the process of new system development in “case company name”, are there any policies or practices that guide the developers regarding topics such as ethics, fairness, accountability?

5. If a software (AI-based solution) causes some harm to the end user / third party, how is accountability distributed between developers and users?

6. Is there a risk mitigation plan for software development in “case company name”? How is it carried out?

7. How well the development process is being documented? For example, can particular decisions or actions made during the development process be tracked back to the individuals behind them?

8. In case of unpredictable system outcome (e.g., autonomous public transportation vehicle is hijacked digitally, AVP system error), is there a particular set of actions?

9. What do you think would be a good way to implement ethics into designing algorithmic systems? E.g., tools, practices, methods?

10. How open is the design and development process of the algorithm to clients and customers?

11. Would you say that that participation of civil society organizations in the design of the algorithm is necessary? Why or why not?

12. Do you think that AI/AS should be made understandable/explainable to the end users?

13. How do you understand the following phrase: “algorithmic audit”?
    Do you think it is necessary to introduce third-party auditors for private companies utilizing algorithmic systems?

14. If public algorithmic auditing standards and compliance policies were to be implemented, what kind of mechanism can be considered to make stakeholders accountable?

15. Have you ever heard of the document called “Ethical Guidelines for AI” by “case company name”?

Table 14. Interview data coding structure

Initial coding	**Selective (focused) coding**	**ACM FAT principles coding**
Describing the link between data and output Indicating the role of data quality in algorithmic bias	Data quality importance	Accuracy
Providing example of constraints related to the black box Referring to algorithm black box Referring to technological constraints of algorithms leading to bias	Technological constraints	
Distinguishing between the area of social and the area of technical Indicating impossibility of connecting ethics and black box algorithms Expressing skepticism regarding AI ethics Linking ethics with aspects outside of the technical area Labelling ethics as a hindrance factor for AI development	AI ethics attitude	Fairness
Recalling a case of algorithmic bias Recalling how algorithmic bias case was solved Referring to algorithmic bias on the basis of demographics Giving an example of non-algorithmic discrimination case based on race Pointing out the illegality of discrimination Distinguishing between algorithmic efficiency and bias	Attitude towards fairness	

Expressing opinion regarding participation of civil society organizations in the design of algorithms Making a claim regarding socio-economic aspect playing a role in algorithmic bias case	Socio-economic aspects in algorithmic bias	
Claiming that making AS explainable is unnecessary	Attitude towards making AS explainable	Explainability
Stating opinion on third-party auditors for algorithms Pointing out issues with legislation	External auditing	Auditability
Describing development process transparency within the company	Transparency	Auditability, explainability
Proposing mechanisms to make stakeholders accountable Making a claim regarding necessity of shared accountability Describing drivers for compliance	Responsibility practices and mechanisms	Responsibility
Pointing out the importance of hierarchy in managing projects Pointing out that only a few responsible employees should be worried about ethics implementation Making a claim that developers should be responsible for algorithmic bias Making a claim that developers should not necessarily be held accountable for algorithmic bias	Internal roles	

Describing ethics-related guidelines and policies Describing risk mitigation policies and actions	Internal policies and guidelines	
Providing the details about the scope of current work 'Ethical Guidelines for AI' awareness Stating an opinion regarding the term 'algorithmic accountability' Stating an opinion regarding the term 'algorithmic audit'	Scope of work and individual documentation/definition awareness	

Appendix B.1-1. Interview summary: respondent ADR/1

Algorithmic accountability project ADR/1	
Interviewee name	Withheld for anonymity purposes
Occupation	Engineer
Interview date	November 3, 2020
Location	Case company, Tokyo headquarters
Duration and data details:	26:57, received permission for audio recording

Interview summary: selected transcript parts

(Disclaimer, interview purpose and scope explanation)

Researcher:
Do you think it is necessary at all for developers (software developers, engineers or anyone who deals with AI/AS) to be able to implement ethics into the systems they develop and to construct them fair by design?

R/ENG:
When it comes to necessity, I think it is a little bit tricky, right? I think that the bottom line is that the companies, they just don't want to get sued, they don't want to do something that will get them in legal trouble. I think they always will try to stay as legal as possible and as cheap as effective as possible; you know? And of course, there is a lot of conflict, so you can think of self-driving cars, like there is a point when a car detects when a collision is going to happen to save a passenger's life rather than a person who is just walking around. This kind of stuff, you can think about ethics and necessity of it, but the bottom line is that they just don't want to get sued and don't want to get accountability. I definitely think it is very important, but at the same time, we got to be careful because it is something very hard to control, like development of AI in general, like any kind of big system really, is very organic and if you just try to hold it down and keep it on a leash to make sure it is always ethical, «perfectly» ethical, I feel like it is going to push it back, you know? Definitely, to a certain extent, pushing the boundaries of it is important, but having too much ethics to take into account might harm it, hold the development down, this is what I'm saying. Like, in a sense that if you always have to account for some kind of stuff, makes the development hundred times more difficult.

Researcher:
So it should be somehow implemented, but taking it too seriously or pushing it too hard may be harmful?

R/ENG:

Yeah. Ideally, you would not have to take it into account if the laws were perfect, because if a law states that the developer is responsible for the kind of damage that they cause, they are not going to push for this kind of stuff, you know? Because they don't want to account, they don't want to get sued or anything, right? But the thing with this kind of technology is that it is really new and it takes years for the legislation to update and to reflect what is actually written about ethics, you know? So during that gap, for the first few years when there is something new, it is a dangerous situation, because legally you are «okay to go», but ethically it is not something so optimal.

Researcher:
So in general, how do you understand the concept of algorithmic accountability? The phrase itself?

R/ENG:
I have never heard about it, but when you say «algorithmic accountability» I think you are just saying, like, having some sort of an algorithm that deals with any kind of human damage potential, just in a way to minimize the damage and how the developers have the responsibility to do that. I don't know if my understanding is correct, but... And I feel that in some areas it is definitely has to happen, because our lives are so much ruled by very complex systems that we barely have any control over.

Researcher:
So you think that it is inevitable in some way?

R/ENG:
Oh yeah, of course. So, let me give you an example. Youtube, right? Youtube has… the only way to actually publish a video anytime, however you want is because they actually have machine learning to filter and any inappropriate videos, such as pornography, violence against animals and other humans - it gets banned automatically. But a lot of times they have to review the data manually. And the problem is that they usually outsource it to countries that are quite conservative, such as Pakistan, places like that, and those manual reviewers will do the review.

Researcher:
I see. I actually read an article before about how this Facebook moderating thing works and people get psychologically damaged, but I have never actually heard that they usually hire moderators from developing, conservative countries.

R/ENG:
Yeah, well, it depends. They usually look for cheap labor.

Researcher:
Makes sense.

R/ENG:
But what happens is, many of these moderators end up banning a lot of videos, such as LGBTQ-related topics and they get a lot of backlash for that, because, you know, machine learning is going to be based on data that is manually reviewed and the algorithm is going to be biased. So, there is definitely a bias that is inhered in humans and it is going to cross over to the machine learning and to algorithms. And in that sense, the developers definitely have some sort of responsibility for this kind of thing.

Researcher:

Let's imagine, in a case when some sort of algorithmic system is utilized in a particular organization is found to produce some faulty outcomes (e.g., bias), who do you think should be held accountable? For example, a third-party company is hired by a public agency to develop an algorithm for justice system. However, after it was developed and widely applied, an independent investigative company ran an analysis and found out that the algorithm made systematic racially discriminating mistakes (it is biased). Who should be held accountable in this case?

R/ENG:
This is a good question. I guess it depends a lot on the case, but developers are not necessarily to be held accountable. A lot of times, you won't develop something fresh, you are just going to base it off data that we have, so I think that... there was a famous case with Google hiring, right? That was based on hiring professionals and they mainly hired men, and the algorithm was biased towards men and that was not something they wanted to do, to happen. So, in that case, it is a difficult question. Because the developers, they are just going to maximize, make the system more effective, I would not account them necessarily for everything that happens, you know? I is hard to say... I honestly cannot say who should be held accountable.

Researcher:
Should it be a shared accountability?

R/ENG:
Yeah, I guess because you cannot directly just sue for damage when that happens, right? Because you are definitely getting some unfair treatment.

Researcher:
Don't you think that it is the developers who should have thought about the potential of bias being introduced into an algorithm? Should they have checked somehow, run some tests? Because bias can find its way into an algorithm through many ways.

R/ENG:
Yeah. So, the difference is, when you are hiring people, and you are using AI, and AI is biased against women for instance, it is very easy to prove, because it a machine, you know? We can analyze the data. But when humans do it, it is much harder to prove. How are they going to prove that the interviewer was biased against women, right? So, I feel like if we allow this kind of stuff to happen with humans, if it so hard for people who get rejected to prove that they were biased against, if we put different standards for AI, it is going to be hypocritical. But I definitely see that with AI it is so much easier. But I see your point, that developers should have the forethought of doing that stuff and not release it into the wild. It is definitely something really new, so we have to think about it.

Researcher:
But then again, if the training data is biased, as the data usually reflects the society on the whole, then the algorithm will be biased too. So maybe that public agency that hired a third-party company should have been accountable for providing this kind of data?

R/ENG:
Yeah, they should be sued. For the damage to public good, this kind of thing. Cross-sectional

lawsuit, definitely.

Researcher:
For example, in «Case company name», how well the development process is being documented? For example, can particular decisions or actions made during the development process be tracked back to the individuals behind them?

R/ENG:
I feel like it depends on the project, but they mostly can, especially what we do with «Project name». We use Git. So, the basic idea is that if *you* change something, it is your name, something very well written, exactly who did what kind of changes, you know. But after certain point, humans commit mistakes, so if the system is faulty, it is also the company's responsibility to have made sure that it was foolproof and tested before. But I feel like to a certain extent it is fairly easy in general to know who did something wrong, just because, you know, they keep track of it.

Researcher:
Do you have any idea about what would be a feasible way to implement ethics into designing algorithmic systems? E.g., tools, practices, methods?

R/ENG:
I think it depends a lot on what the objective is. But we should have a general guideline to minimize any potential sort of damage. From the developers' side, it is something very difficult to implement, so I guess you just have to have some general guideline that you can follow through and of course legislation, like ideally it would not take so many years to update and enforce. Because when you develop something new, you have absolutely no control over it and until the legislation is updated, it is possible to do whatever you want basically. So, it happens a lot and especially because most politicians are old and have no idea what is going on. I feel like ideally there should be something like a ministry specifically with people who are more tech-minded handling this kind of things.

Researcher:
In case of some unpredictable system outcome (e.g., autonomous public transportation vehicle is hijacked digitally, AVP system error), is there a particular set of actions?

R/ENG:
Well, all of these systems, they already have a bunch of fail-savers in place. So, when someone breaks into your home, you are not going to call a constructor and say "Oh, you built my house and it was not foolproof, someone could get in!" The damage in this case is from the people who did the aggression. And it doesn't mean you don't lock your house, right, you still going to lock it. Of course, some level of security is very important, but after a certain point, the fault is on the person who actually committed the act, it is not company responsibility or anything.

Researcher:
Would you say that that participation of civil society organizations in the design of the algorithm is necessary? Why or why not?

R/ENG:
I think it a really good question. Like I said, technology is really about how we use it, so open

discussion is really important. But at the same time, you don't want to bounder it all the time. So, I feel like it is half and half, half of it has to be approved by developers, but you can't just rely on people to be good all the time, so there is definitely some responsibility to push in. We should hold them accountable, and we should have civil society organizations and also government to put precautions in place to protect the general public.

Researcher:
Do you think that AI/AS should be made understandable/explainable to the end users?

R/ENG:
Not necessarily. It is unfortunate, but we live in a world that is so dependent on technology, but we are completely technologically illiterate, we don't know what is going on. So, if we try to "dumb it down", it is just a backward process, it takes too much effort. But there is a lot of misconceptions, yes it would be good to let consumers know what is going on, but most of the time they would not care. And if they do care, most likely they are illiterate in a technological sense, you know. So I do not know how to good about that.

Researcher:
So in general you would say that it is not necessary? It is challenging to implement?

R/ENG:
Yes. It is very challenging for sure.

Researcher:
Is there some kind of solution for this? For people who do not happen to have enough technical literacy, for them to provide some kind of explanation about how the system works?

R/ENG:
Is is complicated. For EU, for example, if someone uses their public data, they have to tell you what and how they are using it. Once again, I feel like legislation is really important, to keep it up after technological development. But educating people about it, it gets very complicated. It feels like a fruitless effort, because there is just too much information.

Researcher:
How do you understand the following phrase: "algorithmic audit"?

R/ENG:
I understand it as a process of analyzing algorithm in a sense of judging it, I guess, its fairness and outcomes, this kind of thing.

Researcher:
Do you think it is necessary to introduce third-party auditors for private companies utilizing algorithmic systems?

R/ENG:
I think not always, but in some cases, for sure. If all public transportation turns to AI-managed system, definitely, we need a third party to make sure that it will not make any dumb mistakes or hurt people.

Researcher:
How do you think this should work? Should it be a government agency, a consulting company, etc.?

R/ENG:
I feel like it should be definitely a government agency, removed from bias and profits. Because it is a public good, we all want technology and all want efficient AI, but it should be completely dissociated with politics, we want people who are technically minded doing this kind of stuff.

Algorithmic accountability project ADR/2	
Interviewee name	Withheld for anonymity purposes
Occupation	R&D
Interview date	November 12, 2020
Location	Case company, Tokyo headquarters
Duration and data details:	43:37, received permission for audio recording

Interview summary: selected transcript parts

(Disclaimer, interview purpose and scope explanation)

Researcher:
Do you think that it is necessary for developers to implement ethics into the systems they develop and to make the systems they develop "fair" by design?

P/R&D:
In that case we need to define what fair is. What do you mean by fair?

Researcher:
So, they have to consider all the options regarding sensitivity of this particular field within which the algorithm will be applied, the data it will deal with, so when it actually is applied all the precautions are taken to minimize the risks regarding the biases that may occur.

P/R&D:
I think I understand what you are talking about. Let me give you an example of this. So there was a time HP made a face detection algorithm and that algorithm at first did not work properly for Black people. And when they did a research about why this happened, because a lot of people were coming out and calling HP racist, but the only reason why it wasn't detecting Black people is because the way we detect faces using structures and it depends on contrast, light of the image. With Black people, what happened was there was a washed-out contrast. So, when you put something like an age detection algorithm on them, it couldn't find ages. And when in my opinion, from the outside it might look like the ethics was being violated, because why is this program not acting fairly. But you have to consider that technology, the way we detect faces might work differently in different circumstances. And HP was very fast to deal with this problem using much better algorithm. It was just a matter of getting the algorithm out of the door first. And I don't think that…nobody even thought about ethics, because we did not do anything that would violate this even. Were you talking about stuff like that?

Researcher:
Yes, this included, but this is more about technical constraints. But, for example, for algorithms that are used for hiring there was this case with Google. This is a better example. Data-based faulty outcomes rather than technical constraints.

P/R&D:
Basically, what I think that problem was, getting into Google is very competitive. The thing is, Google doesn't have to hire everybody, they are going to find the best applicants.
For Google, the algorithm that you are talking about, they were hiring many Asian men. The most hired, and the least hired I am not sure. And there is a reason, there is a socio-economic aspect to it as well. You have heard about tiger moms, right? Forcing being the best from the early age. Of course, that is going to end up in the hiring practices as well, when you see a lot of Asian people getting into top managerial positions and IT companies. The CEO of Google, CEO of Microsoft, CEO of PepsiCo. It is less about ethics and more…The guys making the algorithms, they are less concerned about ethics and more concerned about getting the best person for the jobs. And the ethics part happens outside of the technical area and more of a sociological aspect of society. So, in my opinion with Google, it is less about ethics and more about getting the best person for the job. That is what I think.

Researcher:
Because we have to feed the data to an algorithm first and bias can make its way into the algorithm through many ways and basically the data reflects the views of society on the whole. So that is a problem. So that is how the social field that you were talking about before and the technical field are intertwined.

P/R&D:
We do not mean algorithms like that. We do not make exceptions for race, gender, things like that. It is basically a black box, what happens inside we do not know, we do not write the code for that. That code is written by machines by itself. And machines technically would have no ethics, because machines are not a social animal. So, the black box inside, we do not touch them. We just feed the data, we feed it CVs and the output comes as an answer, yes or no, do we hire, or we don't. Even if we want it to change the black box, we would not know how to, because it is beyond our comprehension. We do not know which segment works with the gender, which works race, which works with the socio-economic background, we do not know that. 12:45 I do not think there is any ethics we can put into this black box. At least that is what I think.

Researcher:
Maybe not about the algorithm itself, in a strictly techno-centric view, but the data this algorithm feeds on. Do you think it is important to consider the age, race, those factors?

P/R&D:
From the legal standpoint, you cannot discriminate on the basis of race, gender. If we make an algorithm that preferentially treats a race or gender, it would be discriminating against other races, etc. Even preferential treatment like this might be considered a violation of equal rights, so that is a problem. For example, do you know the case of Asians in America suing Harvard? What happened is, because there were so many Asian people applying to Harvard, they were like, we have enough Asians, we need diversity. Students for fair admissions versus Harvard, there was lawsuit. And they are saying the Harvard discriminates against Asian Americans.
Of course, we should be mindful about who we hire. We should hire the best people first and then look a gender and other factors later. Because at the end of the day, the company's job is to make money.

Researcher:
So in general you would say it is not necessary to consider making algorithms fair by design, but it is important to make them as technically efficient as possible?

P/R&D:
When you make something fair for everyone, some people are going to be discriminated against. We have people of different heights, for example. Let's say, you walk underneath a bar like this. Ok, let's say, you want to get into a box. You don't want to make a box too big or else small people won't fit into that box properly. But you don't want to make it too small as well, then the tall people won't be able to get into that box. So, if make it medium size, most people would fit in, but the outliers in this case will be in a trouble. If you make things fair by design, some people are going to be better than some people.
Even in tech, the way you want to be fair, the way you want to get more people into tech, is not inspiring to get more people into tech. Inspiring your kids, inspiring your daughters, inspiring other members of society. Because tech is a, I would say, meritocracy.

Researcher:
But on the other hand, some biases within society, they can be actually tracked through application of algorithms, it can go both ways. Some biases that we may not think about and realize that they do exist.

P/R&D:
Algorithms do not have the same biases as humans do. Human bias is much more social, but algorithmic bias is much more numerical. Algorithm wouldn't be biased against, let's say, against Asian person if the data provided to it wasn't already biased and the data provided to it comes from society in general, not the algorithm.

Researcher:
Yes, and the problem is how should accountability be shared in this case?

P/R&D:
The algorithm is not accountable, the algorithm just does the stuff data told it to do. Let's say, I created an algorithm for the C company. I know that all other companies surrounding this company are not hiring Asians. The data that I use, most of the Asians won't be hired, because most of the Asians are not hired and I feed that data to my algorithm. Is the algorithm responsible for nit hiring Asians even though the data provided is biased?

Researcher:
This is what I am talking about. So should the developer, so the people who create this algorithm, consider that this field is problematic, sensitive and these factors should be considered and that is how we should construct this algorithm, with this kind of problem in mind. This is how ethics should be applied in this case?

P/R&D:
Basically, is there is a problem, in my opinion, I would rather deal with the data first, try to unbiased the data instead of changing the algorithm. It is less of a problem with the algorithm and more about the problem with the data it feeds on. More of a social problem rather than technological.

Appendix B.1-3. Interview summary: respondent ADR/3

Algorithmic accountability project ADR/3	
Interviewee name	Withheld for anonymity purposes
Occupation	Engineer
Interview date	November 10, 2020
Location	Case company, Tokyo headquarters
Duration and data details:	45:01, received permission for audio recording

Interview summary

For anonymity purpose the interview transcript is not disclosed. The participant is one of the leading engineers with more than 30 years of experience in the industry (primarily automotive). Discussion was case company specific (internal practices) and respondent addressed automotive industry regulations (CMMI, Automotive Spice), not allowing failure conditions, differences between software development standards for European OEMs and Japanese OEMs, ethical implications and fairness issues among other topics.

Appendix B.1-4. Interview summary: respondent ADR/4

Algorithmic accountability project ADR/4	
Interviewee name	Withheld for anonymity purposes
Occupation	Engineer
Interview date	November 3, 2020
Location	Case company, Tokyo headquarters
Duration and data details:	21:27, received permission for audio recording

Interview summary

For anonymity purposes full interview transcript is not disclosed. The participant belongs to an engineering department and addressed the issue of transparency and fairness in software development, case company specific guidelines and practices, security and ethics-related corporate training, design and development process of the algorithmic systems and compliance mechanisms among other topics.

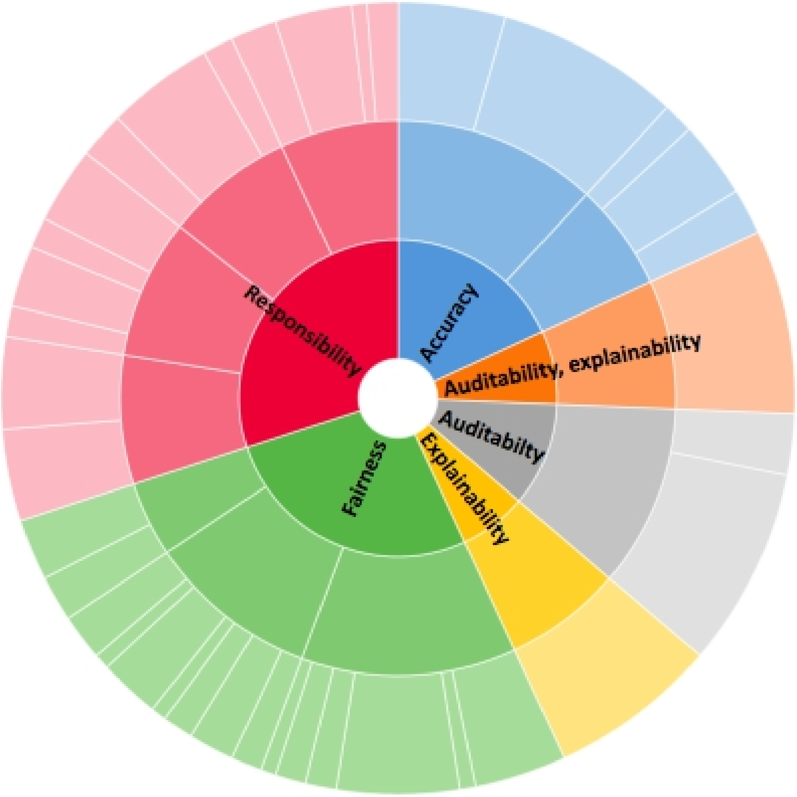
Appendix B.1-5. Interview summary: respondent ADR/5

Algorithmic accountability project ADR/5	
Interviewee name	Withheld for anonymity purposes
Occupation	Engineer
Interview date	July 25, 2020
Location	Case company, Tokyo headquarters

Interview summary

For anonymity purposes full interview transcript is not disclosed. The participant belongs to an engineering department and currently is a computer science major in one of the institutions known to be a center of excellence for Artificial Intelligence research worldwide. The respondent addressed the issues of risk mitigation in software development, participation of civil society organizations in the design of the algorithmic systems, significance and possibility of making algorithmic systems explainable to the public, ethical and social implications of using algorithms among other topics.

Figure 13. Interview coding references: NVivo query



Third layer codes (ACM) compared by number of coding references: NVivo query based on the interview data

Appendix C. Alan Turing Institute, Ben Scnhneiderman lecture on algorithmic accountability summary

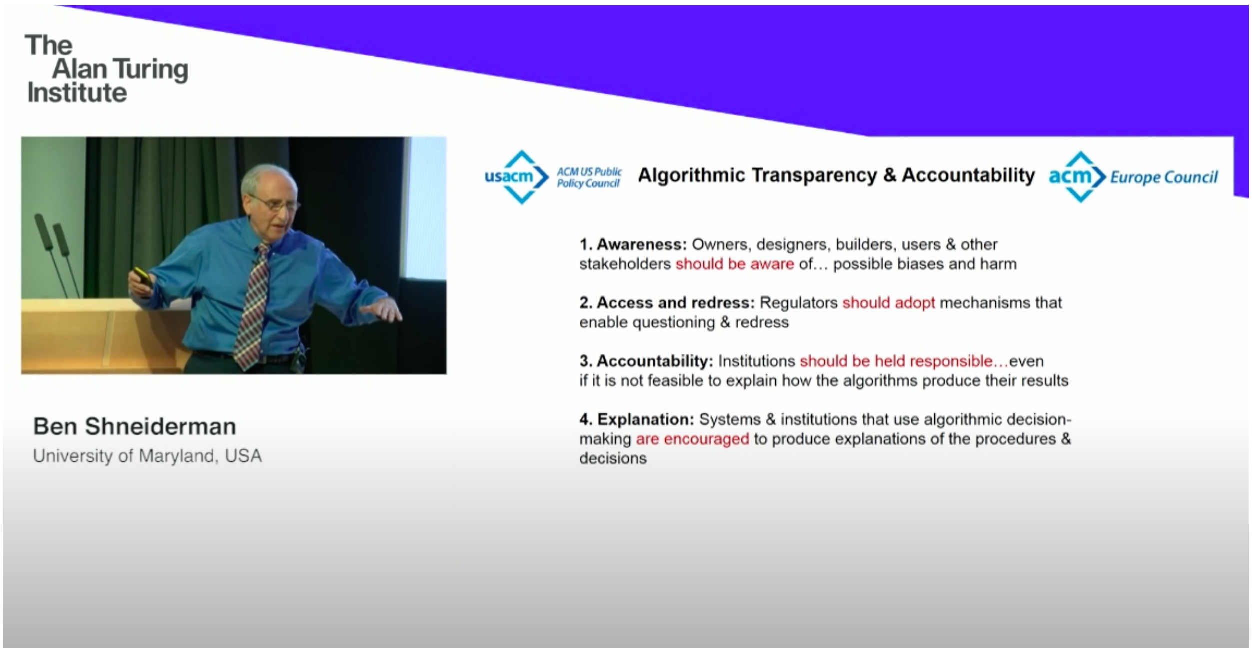
Materials from Ben Scnhneiderman's lecture on algorithmic accountability (*The Alan Turing Institute: Algorithmic Accountability: Professor Ben Shneiderman, University of Maryland - YouTube*, n.d.) were used for stakeholder engagement within Design and Implementation stages.

«Responsibility is the guide to clarifying the design of systems».
If you ensure that the operator or the managers above have the responsibility, then you are doing better. But the current designs do not have this feature.

Human operator has the responsibility, but the machine does not.
Many software contracts still have clauses that state «hold harmless» - the designers, the managers, the human operators will be held harmless, while software is delivered «as is».

«Statement of Algorithmic Transparency and Accountability» by ACM US Public Policy Council (January 12, 2017)

Ben Schneiderman critics of the statement - «should be», «are encouraged» formulations are too vague.



Ben Shneiderman
University of Maryland, USA

Algorithmic Transparency & Accountability

1. **Awareness:** Owners, designers, builders, users & other stakeholders should be aware of… possible biases and harm

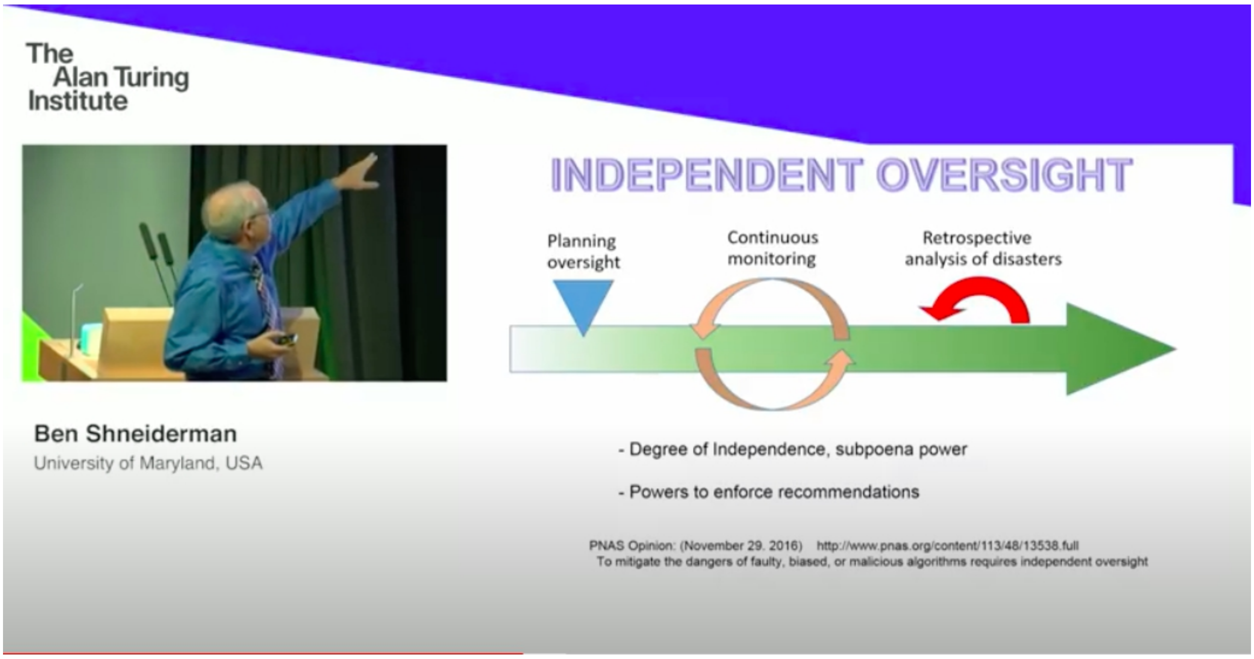
2. **Access and redress:** Regulators should adopt mechanisms that enable questioning & redress

3. **Accountability:** Institutions should be held responsible…even if it is not feasible to explain how the algorithms produce their results

4. **Explanation:** Systems & institutions that use algorithmic decision-making are encouraged to produce explanations of the procedures & decisions

Ensuring human control while increasing automation - there's different kinds of human of control. **Control at multiple levels of an organization**
Independent oversight - proposes to introduce National Algorithms Safety Board (similar to National Transport Safety Board, etc.) (33:00)
Components: Planning Oversight, Continuous Monitoring (expensive, but dramatically effective at reducing violations of algorithms), Retrospective Analysis
Insurance companies will be strong advocates for establishing an investigating body for algorithm safety
To build sympathy in the computing field, we need to show that actually explainable AI is possible

How to ensure control at different layers/levels?
Besides legal solutions (liability), is incentivization of responsibility possible?



The Alan Turing Institute

Ben Shneiderman
University of Maryland, USA

INDEPENDENT OVERSIGHT

- Degree of Independence, subpoena power
- Powers to enforce recommendations

PNAS Opinion: (November 29. 2016) http://www.pnas.org/content/113/48/13538.full
To mitigate the dangers of faulty, biased, or malicious algorithms requires independent oversight

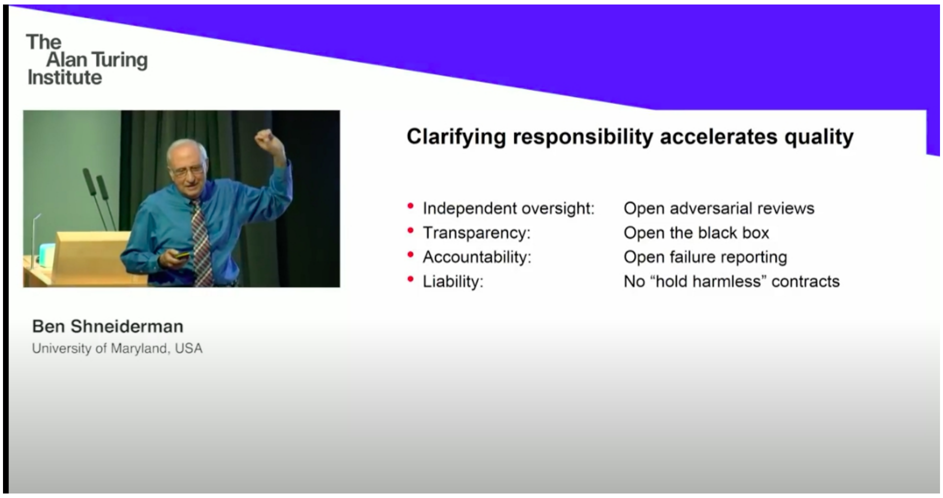
Possible issues: people involved in the oversight should have enough knowledge about the topic, but you don't want to become so close to the people so that they become too friendly and sympathetic

Clarifying responsibility accelerates quality

If we focus on responsibility for failures, then we will see better how to design so that the human - **users, their managers and supervisors going up the chain can actually say they are responsible for the actions.**

Ben Schneiderman suggests that to ensure accountability, we need to open failure reporting.

Air transport system is very safe, because it is very open. Reporting of errors produces a culture of safety. (40:29)



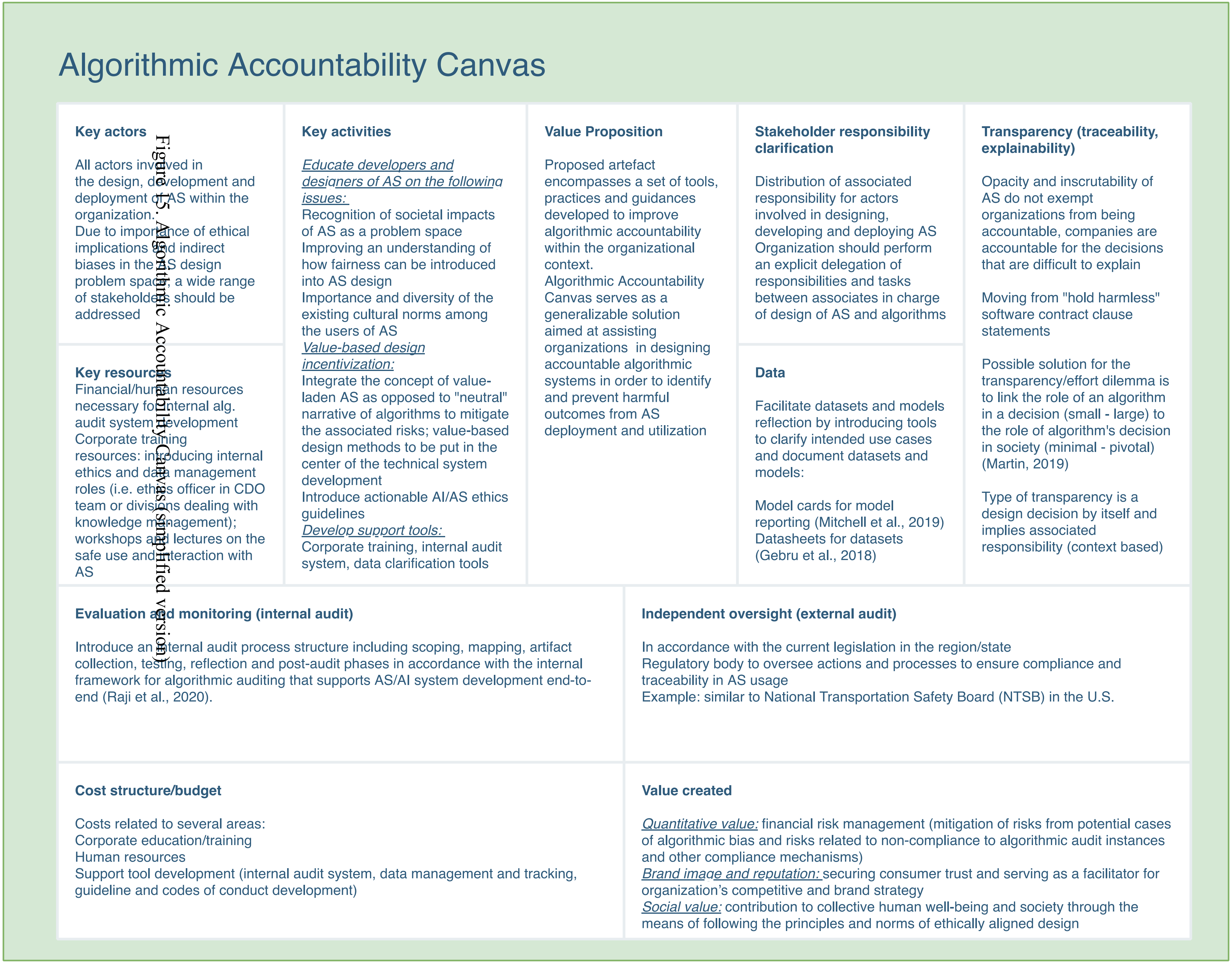
The Alan Turing Institute

Ben Shneiderman
University of Maryland, USA

Clarifying responsibility accelerates quality

• Independent oversight:	Open adversarial reviews
• Transparency:	Open the black box
• Accountability:	Open failure reporting
• Liability:	No “hold harmless” contracts

Figure 14. Algorithmic Accountability Canvas



Algorithmic Accountability Canvas

Key actors

All actors involved in the design, development and deployment of AS within the organization.
Due to importance of ethical implications and indirect biases in the AS design problem space, a wide range of stakeholders should be addressed

Key resources
Financial/human resources necessary for internal alg. audit system development
Corporate training resources: introducing internal ethics and data management roles (i.e. ethics officer in CDO team or divisions dealing with knowledge management); workshops and lectures on the safe use and interaction with AS

Key activities

Educate developers and designers of AS on the following issues:
Recognition of societal impacts of AS as a problem space
Improving an understanding of how fairness can be introduced into AS design
Importance and diversity of the existing cultural norms among the users of AS
Value-based design incentivization:
Integrate the concept of value-laden AS as opposed to "neutral" narrative of algorithms to mitigate the associated risks; value-based design methods to be put in the center of the technical system development
Introduce actionable AI/AS ethics guidelines
Develop support tools:
Corporate training, internal audit system, data clarification tools

Value Proposition

Proposed artefact encompasses a set of tools, practices and guidances developed to improve algorithmic accountability within the organizational context.
Algorithmic Accountability Canvas serves as a generalizable solution aimed at assisting organizations in designing accountable algorithmic systems in order to identify and prevent harmful outcomes from AS deployment and utilization

Stakeholder responsibility clarification

Distribution of associated responsibility for actors involved in designing, developing and deploying AS
Organization should perform an explicit delegation of responsibilities and tasks between associates in charge of design of AS and algorithms

Data

Facilitate datasets and models reflection by introducing tools to clarify intended use cases and document datasets and models:

Model cards for model reporting (Mitchell et al., 2019)
Datasheets for datasets (Gebru et al., 2018)

Transparency (traceability, explainability)

Opacity and inscrutability of AS do not exempt organizations from being accountable, companies are accountable for the decisions that are difficult to explain

Moving from "hold harmless" software contract clause statements

Possible solution for the transparency/effort dilemma is to link the role of an algorithm in a decision (small - large) to the role of algorithm's decision in society (minimal - pivotal) (Martin, 2019)

Type of transparency is a design decision by itself and implies associated responsibility (context based)

Evaluation and monitoring (internal audit)

Introduce an internal audit process structure including scoping, mapping, artifact collection, testing, reflection and post-audit phases in accordance with the internal framework for algorithmic auditing that supports AS/AI system development end-to-end (Raji et al., 2020).

Independent oversight (external audit)

In accordance with the current legislation in the region/state
Regulatory body to oversee actions and processes to ensure compliance and traceability in AS usage
Example: similar to National Transportation Safety Board (NTSB) in the U.S.

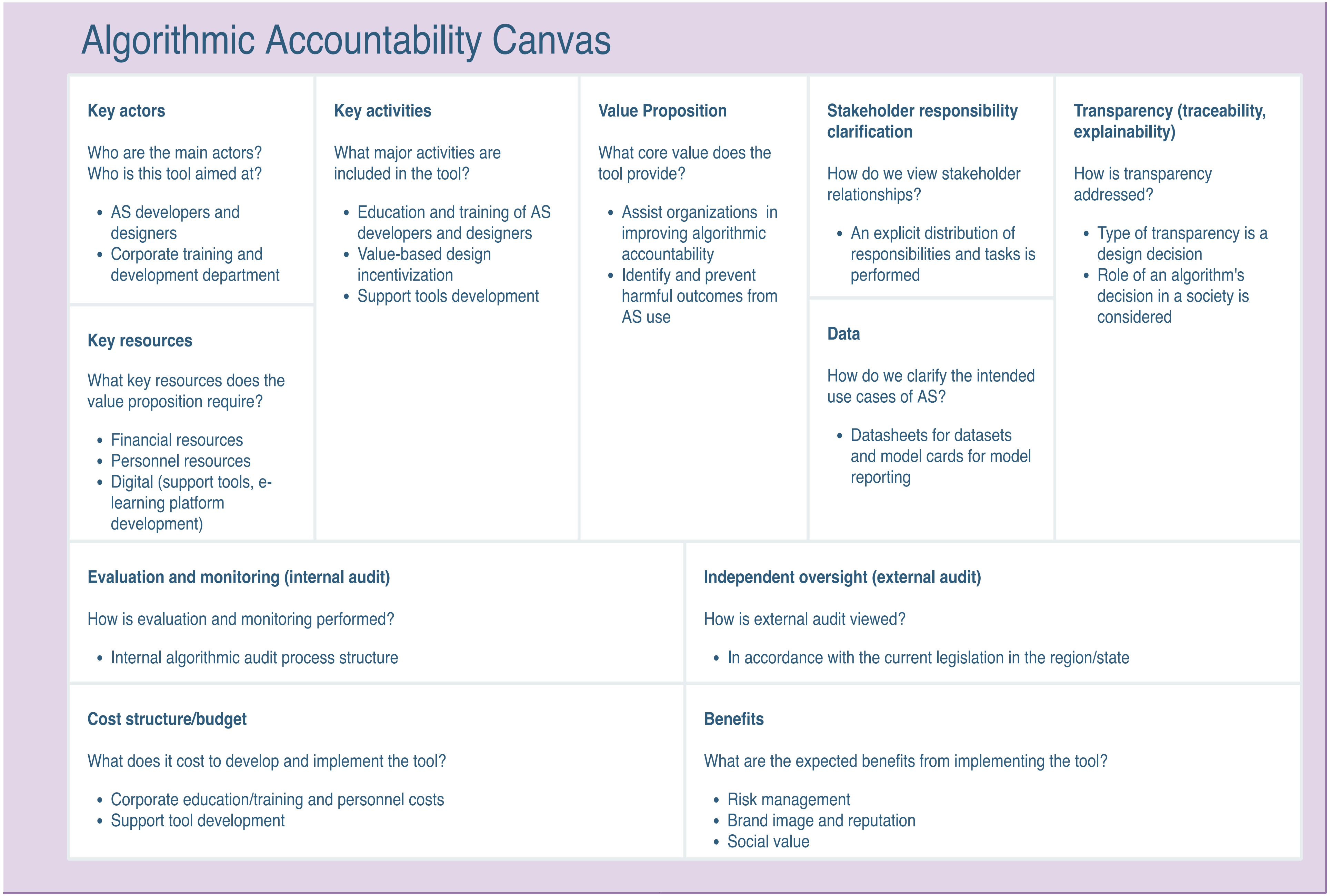
Cost structure/budget

Costs related to several areas:
Corporate education/training
Human resources
Support tool development (internal audit system, data management and tracking, guideline and codes of conduct development)

Value created

Quantitative value: financial risk management (mitigation of risks from potential cases of algorithmic bias and risks related to non-compliance to algorithmic audit instances and other compliance mechanisms)
Brand image and reputation: securing consumer trust and serving as a facilitator for organization's competitive and brand strategy
Social value: contribution to collective human well-being and society through the means of following the principles and norms of ethically aligned design

Figure 15. Algorithmic Accountability Canvas (simplified version)



Algorithmic Accountability Canvas

Key actors

Who are the main actors?
Who is this tool aimed at?

- AS developers and designers
- Corporate training and development department

Key resources

What key resources does the value proposition require?

- Financial resources
- Personnel resources
- Digital (support tools, e-learning platform development)

Key activities

What major activities are included in the tool?

- Education and training of AS developers and designers
- Value-based design incentivization
- Support tools development

Value Proposition

What core value does the tool provide?

- Assist organizations in improving algorithmic accountability
- Identify and prevent harmful outcomes from AS use

Stakeholder responsibility clarification

How do we view stakeholder relationships?

- An explicit distribution of responsibilities and tasks is performed

Data

How do we clarify the intended use cases of AS?

- Datasheets for datasets and model cards for model reporting

Transparency (traceability, explainability)

How is transparency addressed?

- Type of transparency is a design decision
- Role of an algorithm's decision in a society is considered

Evaluation and monitoring (internal audit)

How is evaluation and monitoring performed?

- Internal algorithmic audit process structure

Independent oversight (external audit)

How is external audit viewed?

- In accordance with the current legislation in the region/state

Cost structure/budget

What does it cost to develop and implement the tool?

- Corporate education/training and personnel costs
- Support tool development

Benefits

What are the expected benefits from implementing the tool?

- Risk management
- Brand image and reputation
- Social value

References

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society*. https://doi.org/10.1177/1461444816676645

Anderson, R. E. (1992). ACM Code of Ethics and Professional Conduct. *Communications of the ACM*. https://doi.org/10.1145/129875.129885

Association for Computing Machinery US Public Policy Council (USACM). (2017). *Statement on algorithmic transparency and accountability*. USACM Press Releases.

Awarding GCSE, AS & A levels in summer 2020: interim report - GOV.UK. (n.d.). Retrieved March 19, 2021, from https://www.gov.uk/government/publications/awarding-gcse-as-a-levels-in-summer-2020-interim-report

Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*. https://doi.org/10.1016/j.intcom.2010.07.003

Binns, R. (2018). Algorithmic Accountability and Public Reason. *Philosophy and Technology*, *31*(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

Blodgett, J. G., Lu, L. C., Rose, G. M., & Vitell, S. J. (2001). Ethical sensitivity to stakeholder interests: A cross-cultural comparison. In *Journal of the Academy of Marketing Science*. https://doi.org/10.1177/03079459994551

Blum, B. (1992). *Software Engineering:A Holistic View*. Oxford University Press.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework1. *European Law Journal*. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Bovens, M., Goodin, R. E., Schillemans, T., Bovens, M., Schillemans, T., & Goodin, R. E. (2014). Public Accountability. In *The Oxford Handbook of Public Accountability*. https://doi.org/10.1093/oxfordhb/9780199641253.013.0012

Boyce, C., & Neale, P. (2006). Conducting In-Depth Interviews: A Guide for Designing and Conducting In-Depth Interviews for Evaluation Input. *Pathfinder International*.

Brown, A., Chouldechova, A., Putnam-Hornstein, E., Tobin, A., & Vaithianathan, R. (2019). Toward algorithmic accountability in public services a qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300271

Bryant, A., & Charmaz, K. (2012). The SAGE Handbook of Grounded Theory. In *The SAGE Handbook of Grounded Theory*. https://doi.org/10.4135/9781848607941

Buhmann, A., Paßmann, J., & Fieseler, C. (2020). Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse. *Journal of Business Ethics*, *163*(2), 265–280. https://doi.org/10.1007/s10551-019-04226-4

Carayon, P. (2006). Human factors of complex sociotechnical systems. *Applied Ergonomics*.

https://doi.org/10.1016/j.apergo.2006.04.011

Chan, A. W. H., & Cheung, H. Y. (2012). Cultural Dimensions, Ethical Sensitivity, and Corporate Governance. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-011-1146-9

Chandra, L., Seidel, S., & Gregor, S. (2015). Prescriptive knowledge in IS research: Conceptualizing design principles in terms of materiality, action, and boundary conditions. *Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/10.1109/HICSS.2015.485

Charmaz, K. (2006). Memo-writing. In *Constructing Grounded Theory. A practical guide through qualitative analysis*.

Charmaz, K., & Belgrave, L. L. (2012). Qualitative interviewing and grounded theory analysis. In *The SAGE Handbook of Interview Research: The Complexity of the Craft*. https://doi.org/10.4135/9781452218403.n25

Chung, K. Y., Eichenseher, J. W., & Taniguchi, T. (2008). Ethical perceptions of business students: Differences between East Asia and the USA and among "confucian" cultures. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-007-9391-7

Clarke, Y. D. (2019). *All Info - H.R.2231 - 116th Congress (2019-2020): Algorithmic Accountability Act of 2019*. https://www.congress.gov/bill/116th-congress/house-bill/2231/all-info

Clavell, G. G., Zamorano, M. M. n., Castillo, C., Smith, O., & Matic, A. (2020). Auditing algorithms: On lessons learned and the risks of data minimization. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 265–271. https://doi.org/10.1145/3375627.3375852

Clegg, C. W. (2000). Sociotechnical principles for system design. *Applied Ergonomics*. https://doi.org/10.1016/S0003-6870(00)00009-0

Corts, K. S. (2007). Teams versus individual accountability: Solving multitask problems through job design. *RAND Journal of Economics*. https://doi.org/10.1111/j.1756-2171.2007.tb00078.x

Courtland, R. (2018). Bias detectives: The researchers striving to make algorithms fair news-feature. *Nature*, *558*(7710), 357–360. https://doi.org/10.1038/d41586-018-05469-3

Cummings, L. L., & Anton, R. J. (1990). The logical and appreciative dimensions of accountability. In *Appreciative management and leadership: The power of positive thought and action in organizations*.

Diakopoulos, N. (2013). Sex, Violence, and Autocomplete Algorithms. *Slate*.

Diakopoulos, N. (2014). Algorithmic accountability reporting: On the investigation of black boxes. In *A Tow/Knight Brief*. https://doi.org/10.1038/sj.bjp.0701242

Diakopoulos, N. (2015). Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism*. https://doi.org/10.1080/21670811.2014.976411

Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., Howe, B., Jagadish, H. V., Unsworth, K., Sahuguet, A., Venkatasubramanian, S., Wilson, C., Yu, C., Zevenbergen, B., FAT/ML, Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Hay, M., … Zevenbergen, B. (2018). Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. *Fatml.Org*.

Diakopoulos, N., & Koliska, M. (2017). Algorithmic Transparency in the News Media. *Digital Journalism*. https://doi.org/10.1080/21670811.2016.1208053

Donovan, J., Caplan, R., Matthews, J., & Hanson, L. (2018). Algorithmic accountability: A primer. *Data & Society*, *501*(c).

Dose, J. J., & Klimoski, R. J. (1995). Doing the right thing in the workplace: Responsibility in the face of accountability. *Employee Responsibilities and Rights Journal*. https://doi.org/10.1007/BF02621254

Dremel, C., Stoeckli, E., & Wulf, J. (2020). Management of analytics-as-a-service - results from an action design research project. *Journal of Business Analytics*. https://doi.org/10.1080/2573234X.2020.1740616

Dubnick, M. J. (2002). Seeking Salvation for Accountability. *Annual Meeting of the American Political Science Association*.

Dudin, M. N., Kutsuri, G. N., Fedorova, I. J. evna, Dzusova, S. S., & Namitulina, A. Z. (2015). The innovative business model canvas in the system of effective budgeting. *Asian Social Science*. https://doi.org/10.5539/ass.v11n7p290

Eisenhardt, K. M. (1989). Agency Theory: An Assessment and Review. *Academy of Management Review*. https://doi.org/10.5465/amr.1989.4279003

Eslami, M., Vaccaro, K., Karahalios, K., & Hamilton, K. (2017). "Be careful; Things can be worse than they appear" - Understanding biased algorithms and users' behavior around them in rating platforms. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, *Icwsm*, 62–71.

Fernando, M., & Chowdhury, R. M. M. I. (2010). The relationship between spiritual well-being and ethical orientations in decision making: An empirical study with business executives in Australia. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-009-0355-y

Fox, W. M. (1995). Sociotechnical System Principles and Guidelines: Past and Present. *The Journal of Applied Behavioral Science*. https://doi.org/10.1177/0021886395311009

Frink, D. D., Hall, A. T., Perryman, A. A., Ranft, A. L., Hochwarter, W. A., Ferris, G. R., & Todd Royle, M. (2008). Meso-level theory of accountability in organizations. In *Research in Personnel and Human Resources Management*. https://doi.org/10.1016/S0742-7301(08)27005-2

Frink, D. D., & Klimoski, R. J. (1998). Toward a theory of accountability in organizations and human resource management. In *Research in personnel and human resources management*.

Frink, D. D., & Klimoski, R. J. (2004). Advancing accountability theory and practice: Introduction to the human resource management review special edition. *Human Resource Management Review*. https://doi.org/10.1016/j.hrmr.2004.02.001

Fritscher, B., & Pigneur, Y. (2014). Visualizing business model evolution with the Business Model Canvas: Concept and tool. *Proceedings - 16th IEEE Conference on Business*

Informatics, CBI 2014. https://doi.org/10.1109/CBI.2014.9

Garfinkel, S., Matthews, J., Shapiro, S. S., & Smith, J. M. (2017). Toward algorithmic transparency and accountability. In *Communications of the ACM*. https://doi.org/10.1145/3125780

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2018). Datasheets for datasets. In *arXiv*.

Gillespie, T. (2014). The Relevance of Algorithms. In *Media Technologies*. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Glaser, B. G., & Holton, J. (2007). Remodeling grounded theory. *Historical Social Research*.

Glaser, B., & Strauss, A. (1967). The discovery of grounded theory. 1967. *Weidenfield & Nicolson, London*.

Gregor, S. (2006). The nature of theory in Information Systems. *MIS Quarterly: Management Information Systems*. https://doi.org/10.2307/25148742

Gregor, S., Imran, A., & Turner, T. (2014). A “sweet spot” change strategy for a least developed country: Leveraging e-Government in Bangladesh. *European Journal of Information Systems*. https://doi.org/10.1057/ejis.2013.14

Grytz, R., Krohn-Grimberghe, A., & Müller, O. (2020). Business Intelligence & Analytics Cost Accounting: an Action Design Research Approach. *Proceedings of the 28th European Conference on Information Systems (ECIS), An Online AIS Conference*.

Haj-Bolouri, A. (2019). Design Principles for E-Learning that Support Integration Work: A Case of Action Design Research. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-19504-5_20

Hall, A. T., Frink, D. D., & Buckley, M. R. (2017). An accountability account: A review and synthesis of the theoretical and empirical research on felt accountability. *Journal of Organizational Behavior*. https://doi.org/10.1002/job.2052

Herrmann, T., Loser, K. U., & Jahnke, I. (2007). Sociotechnical walkthrough: A means for knowledge integration. *Learning Organization*. https://doi.org/10.1108/09696470710762664

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, *19*(2), 87–92.

Hevner, A. R., March, S. T., Park, J., & Sudha, R. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75–105.

Hochwarter, W. A., Perrewé, P. L., Hall, A. T., & Ferris, G. R. (2005). Negative affectivity as a moderator of the form and magnitude of the relationship between felt accountability and job tension. *Journal of Organizational Behavior*. https://doi.org/10.1002/job.324

Husted, B. W., & Allen, D. B. (2006). Corporate social responsibility in the multinational enterprise: Strategic and institutional approaches. *Journal of International Business Studies*. https://doi.org/10.1057/palgrave.jibs.8400227

Katell, M., Young, M., Dailey, D., Herman, B., Guetler, V., Tam, A., Binz, C., Raz, D., &

Krafft, P. M. (2020). Toward situated interventions for algorithmic equity: Lessons from the field. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 45–55. https://doi.org/10.1145/3351095.3372874

Katz, D., & Kahn, R. L. (1978). The Social Psychology of Organizations (Chapter 13). In *The Social Psychology of Organizations*.

Keijzer-Broers, W. J. W., & de Reuver, M. (2016). Applying agile design sprint methods in action design research: Prototyping a health and wellbeing platform. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-39294-3_5

Kemper, J., & Kolkman, D. (2019). Transparent to whom? No algorithmic accountability without a critical audience. *Information Communication and Society*. https://doi.org/10.1080/1369118X.2018.1477967

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information Communication and Society*. https://doi.org/10.1080/1369118X.2016.1154087

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/2858036.2858402

Kowalski, R. (1979). Algorithm = Logic + Control. *Communications of the ACM*. https://doi.org/10.1145/359131.359136

Kroon, M. B. R., Hart, P., & van Kreveld, D. (1991). Managing group decision making processes: Individual versus collective accountability and groupthink. In *International Journal of Conflict Management*. https://doi.org/10.1108/eb022695

Laird, M. D., Harvey, P., & Lancaster, J. (2015). Accountability, entitlement, tenure, and satisfaction in Generation Y. *Journal of Managerial Psychology*. https://doi.org/10.1108/JMP-08-2014-0227

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*.

Lee, A. S., & Baskerville, R. L. (2003). Generalizing Generalizability in Information Systems Research. *Information Systems Research*. https://doi.org/10.1287/isre.14.3.221.16560

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-Making Processes. *Philosophy & Technology*, *31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. In *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.125.2.255

Mackey, J. D., Brees, J. R., McAllister, C. P., Zorn, M. L., Martinko, M. J., & Harvey, P. (2018). Victim and Culprit? The Effects of Entitlement and Felt Accountability on Perceptions of Abusive Supervision and Perpetration of Workplace Bullying. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-016-3348-7

Marciszewski, W. (1981). Dictionary of logic as applied in the study of language : concepts, methods, theories. In *Nijhoff international philosophy series*.

Markus, M. L., Majchrzak, A., & Gasser, L. (2002). A design theory for systems that support emergent knowledge processes. *MIS Quarterly: Management Information Systems*.

Martin, K. (2019). Ethical Implications and Accountability of Algorithms. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-018-3921-3

Matthew, F. (2008). Software Studies - A Lexicon. In *The MIT Press*. https://doi.org/10.7551/mitpress/9780262062749.001.0001

Maurya, A. (2012). Lean Canvas. *Running Lean Plan That Works*.

Maurya, A. (2014). Why Lean Canvas vs Business Model Canvas ? *Running Lean*.

McNamara, A., Smith, J., & Murphy-Hill, E. (2018). Does ACM's code of ethics change ethical decision making in software development? *ESEC/FSE 2018 - Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 729–733. https://doi.org/10.1145/3236024.3264833

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287596

Miyazaki, S. (2012). Algorhythmics: Understanding Micro-Temporality in Computational Cultures. *Computational Culture*.

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. In *arXiv*.

Muhtaroglu, F. C. P., Demir, S., Obali, M., & Girgin, C. (2013). Business model canvas perspective on big data applications. *Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013*. https://doi.org/10.1109/BigData.2013.6691684

Mullarkey, M. T., & Hevner, A. R. (2015). Entering action design research. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-18714-3_8

Mullarkey, M. T., & Hevner, A. R. (2019). An elaborated action design research process model. *European Journal of Information Systems*. https://doi.org/10.1080/0960085X.2018.1451811

Myers, M. D. (2009). *Qualitative Research in Business & Management*. SAGE.

Myers, M. D., & Venable, J. R. (2014). A set of ethical principles for design science research in information systems. *Information and Management*, *51*(6), 801–809. https://doi.org/10.1016/j.im.2014.01.002

New, J., & Castro, D. (2018). *How Policymakers Can Foster Algorithmic Accountability*. http://www2.datainnovation.org/2018-algorithmic-accountability.pdf

Neyland, D. (2016). Bearing Account-able Witness to the Ethical Algorithmic System. *Science, Technology, & Human Values*, *41*(1), 50–76. https://doi.org/10.1177/0162243915598056

Niemi, E., & Laine, S. (2016). Competence management system design principles: Action design research. *2016 International Conference on Information Systems, ICIS 2016*.

Nunamaker, J. F., Briggs, R. O., Derrick, D. C., & Schwabe, G. (2015). The Last Research Mile: Achieving Both Rigor and Relevance in Information Systems Research. *Journal of*

Management Information Systems, *32*(3), 10–47. https://doi.org/10.1080/07421222.2015.1094961

Ojasalo, J., & Ojasalo, K. (2018). Service Logic Business Model Canvas. *Journal of Research in Marketing and Entrepreneurship*. https://doi.org/10.1108/JRME-06-2016-0015

Osterwalder, A. (2004). The business model ontology a proposition in a design science approach. *Doctoral Dissertation, Université de Lausanne, Faculté Des Hautes Études Commerciales*.

Osterwalder, Alexander, & Pigneur, Y. (2010a). *Business Model Generation - Canvas*. Wiley.

Osterwalder, Alexander, & Pigneur, Y. (2010b). Osterwalder, A., & Pigneur, Y. (2010). Business Model Generation - Canvas. In *Wiley*.

Osterwalder, Alexander, Pigneur, Y., Bernarda, G., & Smith, A. (2014). Value Proposition Design - How to Make Stuff People Want. In *Entwickeln Sie Produkte und Services, die Ihre Kunden wirklich wollen. Die Fortsetzung des Bestsellers Business Model Generation!*

Pan, S. L., Li, M., Pee, L. G., & Sandeep, M. S. (2020). Sustainability Design Principles for a Wildlife Management Analytics System: An Action Design Research. *European Journal of Information Systems*. https://doi.org/10.1080/0960085X.2020.1811786

Pasquale, F. (2015). The Black Box Society. In *The Black Box Society*. https://doi.org/10.4159/harvard.9780674736061

Polack, P. (2020). Beyond algorithmic reformism: Forward engineering the designs of algorithmic systems. *Big Data and Society*. https://doi.org/10.1177/2053951720913064

Purao, S., Rossi, M., & Bush, A. (2002). Towards an understanding of the use of problem and design spaces during object-oriented system development. *Information and Organization*. https://doi.org/10.1016/S1471-7727(02)00006-4

Rader, E., Cotter, K., & Cho, J. (2018). Explanations as mechanisms for supporting algorithmic transparency. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3173574.3173677

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Recker, J. (2012). Scientific research in information systems: ethical considerations in research. In *Scientific Research in Information Systems: A Beginner's Guide*.

Reibenspiess, V., Drechsler, K., Eckhardt, A., & Wagner, H. T. (2020). Tapping into the wealth of employees' ideas: Design principles for a digital intrapreneurship platform. *Information and Management*. https://doi.org/10.1016/j.im.2020.103287

Roberts, J. (1989). Aristotle on Responsibility for Action and Character. *Ancient Philosophy*. https://doi.org/10.5840/ancientphil19899123

Ropohl, G. (1999). Philosophy of socio-technical systems. *Techne: Research in Philosophy and Technology*. https://doi.org/10.5840/techne19994311

Royle, M. Todd and Hall, A. T. (2012). The Relationship between McClelland's Theory of Needs, Feeling Individually Accountable, and Informal Accountability for Others. *International Journal of Management and Marketing Research*. https://doi.org/10.5539/gjhs.v4n2p2

Schedler, A. (1999). Conceptualizing Accountability. *The Self-Restraining State: Power and Accountability in New Democracies*.

Schillemans, T. (2013). Moving Beyond The Clash of Interests: On stewardship theory and the relationships between central government departments and public agencies. *Public Management Review*. https://doi.org/10.1080/14719037.2012.691008

Schlenker, B. R. (1986). Self-Identification: Toward an Integration of the Private and Public Self. In *Public Self and Private Self*. https://doi.org/10.1007/978-1-4613-9564-5_2

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The Triangle Model of Responsibility. *Psychological Review*. https://doi.org/10.1037/0033-295x.101.4.632

Schlenker, B. R., & Weigold, M. F. (1992). Interpersonal processes involving impression regulation and management. *Annual Review of Psychology*. https://doi.org/10.1146/annurev.ps.43.020192.001025

Schouten, B., Klerks, G., Den Hollander, M., & Hansen, N. B. (2020). Action Design Research Shaping University-Industry Collaborations for Wicked Problems. *32nd Australian Conference on Human-Computer Interaction (OzCHI'20)*, 1–15.

Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action design research. *MIS Quarterly: Management Information Systems*. https://doi.org/10.2307/23043488

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3287560.3287598

Senabre Hidalgo, E., & Fuster Morell, M. (2019). Co-designed strategic planning and agile project management in academia: case study of an action research group. *Palgrave Communications*. https://doi.org/10.1057/s41599-019-0364-0

Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, *98*(March), 277–284. https://doi.org/10.1016/j.chb.2019.04.019

Simga-Mugan, C., Daly, B. A., Onkal, D., & Kavut, L. (2005). The influence of nationality and gender on ethical sensitivity: An application of the issue-contingent model. *Journal of Business Ethics*. https://doi.org/10.1007/s10551-004-4601-z

Sort, J. C., & Nielsen, C. (2018). Using the business model canvas to improve investment processes. *Journal of Research in Marketing and Entrepreneurship*. https://doi.org/10.1108/JRME-11-2016-0048

Strauss, A. L., & Corbin, J. M. (1990). Grounded theory procedures and techniques. In *Basics of*

Qualitative Research.

Tan, S., Adebayo, J., Inkpen, K., & Kamar, E. (2018). Investigating human + machine complementarity for recidivism predictions. In *arXiv*.

Tetlock, P. E. (1992). The impact of accountability on judgment and choice: Toward a social contingency model. *Advances in Experimental Social Psychology*. https://doi.org/10.1016/S0065-2601(08)60287-7

The Alan Turing Institute: Algorithmic Accountability: Professor Ben Shneiderman, University of Maryland - YouTube. (n.d.). Retrieved January 22, 2021, from https://www.youtube.com/watch?v=UWuDgY8aHmU

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. In *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*.

Vakkuri, V., & Abrahamsson, P. (2018). The Key Concepts of Ethics of Artificial Intelligence. *2018 IEEE International Conference on Engineering, Technology and Innovation, ICE/ITMC 2018 - Proceedings*. https://doi.org/10.1109/ICE.2018.8436265

van Wyk, Q., van Biljon, J., & Schoeman, M. (2020). Knowledge Visualization for Sensemaking: Applying an Elaborated Action Design Research Process in Incident Management Systems. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-030-64823-7_14

Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*. https://doi.org/10.1145/3173574.3174014

Warren, J., Lipkowitz, J., Sokolov, V., Konovalenko, I., Kuznetsova, E., Miller, A., Miller, B., Popov, A., Shepelev, D., Stepanyan, K., Meghoe, A., Loendersloot, R., Tinga, T., Loendersloot, R., Rosas-Arias, L., Portillo-Portillo, J., Hernandez-Suarez, A., Olivares-Mercado, J., Sanchez-Perez, G., … Villalba, G. (2019). Algorithmic Accountability Policy Toolkit. *IEEE Intelligent Transportation Systems Magazine*, *October*. https://doi.org/10.3390/s18093010

Washington, A. (2019). How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *The Colorado Technology Law Journal*.

Watson, D., & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.96.3.465

Watson, H. J., & Nations, C. (2019). Addressing the growing need for algorithmic transparency. *Communications of the Association for Information Systems*. https://doi.org/10.17705/1CAIS.04526

Weng, Y. H., & Hirata, Y. (2018). Ethically Aligned Design for Assistive Robotics. *2018 International Conference on Intelligence and Safety for Robotics, ISR 2018*.

https://doi.org/10.1109/IISR.2018.8535889

Wieringa, M. (2020). *What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability*. 1–18. https://dl.acm.org/doi/abs/10.1145/3351095.3372833

Young, M., Katell, M., & Krafft, P. M. (2019). Municipal surveillance regulation and algorithmic accountability. *Big Data & Society*, *6*(2), 205395171986849. https://doi.org/10.1177/2053951719868492