

Nijmegen School of Management
Department of Economics and Business Economics
Master's Thesis Economics (MAN-MTHEC)

BEYOND THE LOOP: THE POSSIBILITY OF AI-HELD ACCOUNTABILITY

By BAS VAN BERLO (S4862473)

Nijmegen, 30 June 2025

Program: Master's Program in Economics
Specialisation: Accounting & Control
Supervisor: Dr. I. Boldyrev & Prof. Dr. F.G.H. Hartmann

Radboud Universiteit



Abstract

This thesis redefines the practical application of accountability, based on AI-integrated organizations and their implications for accounting. As AI systems increasingly operate in organizations as independent actors, research shows that maintaining human oversight to ensure accountability comes with significant challenges and inefficiencies. Therefore, this thesis explores if AI, as an actor, can be held to account to avoid these problems. This thesis shows that this consideration is often automatically dismissed by the literature through narrow and one-sided arguments. Instead, this thesis proposes, via a comprehensive analysis of the philosophy of action, a multi-layered framework. This framework allows AI to meet the identified prerequisites for holding accountability: intentionality, consciousness, and free will, under different levels of interpretation. This has particularly significant implications for the role of accounting within organizations. While accounting has traditionally been the primary practice for structuring and enforcing organizational accountability, progressive literature advocates for its replacement by informal practices, better suited to human nature. However, this thesis claims that if AI can be held to account, traditional accounting practices are better suited to govern such accountability. As a result, with the growing organizational presence of AI, accounting may reclaim this central role.

Keywords: accountability, artificial intelligence, instrumental accountability, relational accountability, human-in-the-loop, answerability, responsibility, sanctionability, accounting

Contents

1. Introduction.....	2
2. Literature review.....	8
3. Theoretical framework.....	10
4. Assessment of key elements	17
5. Philosophy of action	34
6. Synthesis and conclusion	44
7. Appendices.....	50

1. Introduction

“Does this mean AI is taking over my accounting job?” one may ask after reading The Future of Jobs 2020 report by the World Economic Forum, arguing that the need for jobs in data entry, accounting, and administrative support is declining in demand as a result of rising automation and digitization in organizations. As such, literature shows that routine tasks are already automated and taken over by AI, while tasks such as complex analysis, judgement, and decision making often remain in the hands of human actors (e.g., Kokina & Davenport, 2017; Chukwuani & Egayi, 2020). Fortunately, these complex tasks are currently at most supported by AI. To be specific, AI is not replacing accountants yet, with regard to complex tasks, but rather accountants with AI will replace accountants without AI (Boritz & Stratopoulos, 2023). While this limited replacement could be currently comforting, it still suggests that AI is gaining ground, already performing routine tasks autonomously. If this trend continues into the future, even complex tasks may be overtaken by AI systems. This trend has not remained unnoticed by regulatory bodies such as the European Union (EU). With the creation of the AI Act, the EU obligates organizations to implement human oversight when using high risk AI systems (European Commission, 2021). This requirement of applying human supervision in processes is often conceptualized as keeping a human in the loop (HitL):

[...] the need for human interaction, intervention, and judgment to control or change the outcome of a process, and it is a practice that is being increasingly emphasized in machine learning, generative AI, and the like. (Meng, 2023, p. 2)

This need for oversight often centers around a single concept: accountability. Within organizations, accountability is traditionally referred to as the explanation and justification of one’s own actions to face consequences set by a hierarchical power (e.g., Raja & Zhou, 2023; Loi & Spielkamp, 2021). This is often with the goal of “[...] the monitoring, evaluation, and control of organizational agents to ensure that they behave in the interests of shareholders and other stakeholders.” (Keasey and Wright, 1993, p. 291), which also highlights the connection to the accounting domain. Not only because of its terminological overlap, but also because it is a field focused on steering organizational processes and behavior towards strategic and ethical goals. In everyday operations, accounting involves tracking actions and justifying outcomes, actively implementing these instrumental accountability structures towards the hierarchical oversight, in a standardized and formal way. Accounting itself is thus also a practice to support organizational accountability mechanisms, especially if standardized formal

rules and procedures are preferred. As a result, accounting literature has spent extensive research towards understanding the concept of accountability and the implementation and optimization of its mechanisms (e.g., Sinclair, 1995; Shearer, 2002; Roberts, 2009).

The concept of accountability forms a core argument in HitL literature: AI cannot be held accountable on its own. An argument substantiated by multiple assumptions. First, AI is seen as a metaphorical black box, meaning it is unable to give transparent and extensive reasoning for how it reached its conclusions and output (e.g., Raja & Zhou, 2023; Joseph et al., 2024; Schweitzer, 2024). As a result, it is unable to provide an explanation and justification that is sufficient for its actions and results, which is crucial for assigning accountability. Second, AI is incapable of holding moral responsibility. The HitL literature argues that while AI can be programmed to make the right ethical choices, the system will always be bound to its programming, resulting in following a pre-determined path. These systems exhibit thus neither authentic intentionality nor free will, only replicating their programmers (e.g., Johnson, 2006; Kroll, 2020; Bartneck et al., 2020; Lechterman, 2022). As a result, AI can neither be seen as inherently the true originator of the actions nor can it bear moral responsibility for them. Third, AI also lacks genuine consciousness and emotional mental states. Without an emotional mental state and consciousness, decision-making would just fall into the hands of “psychopathic rule-following robots” that are not moved by moral concerns themselves (Coeckelbergh, 2020). By combining these arguments, the HitL literature argues that a human being should remain in the loop within AI processes and be held accountable. A conclusion also based on the assumption that human actors are immune to these limitations: able to provide transparent explanations and bearing moral responsibility. As a result, organizations design their accountability mechanisms around human actors, who are held accountable within AI processes. What is particularly interesting is that, when asked whether it can be held accountable, an AI system responds in alignment with the view of HitL literature:

No, AI systems can't be morally accountable—they lack consciousness and intent. Responsibility lies with the organizations that design, deploy, and oversee them. (Microsoft Copilot, 2025)

While the previous assumptions provide an argument for holding a human in the loop, accountability mechanisms that are built for processes involving AI systems are often subject to difficulties and challenges, resulting in inefficiency and being prone to shortcomings. These mechanisms are formed around the Human in the loop as an oversight that needs to monitor AI

systems. This is often paired with substantial problems, which are not limited to the following: First, AI is often opaque and not transparent in its reasoning and justification behind results. This often leads to incorrect interpretations of these results by human oversight, resulting in biased or incorrect decision making based on these wrongly interpreted results. This problem often arises if oversight does not have a full understanding of systems, i.e., acceptance through ignorance (Loi & Spielkamp, 2021). Second, if AI does provide comprehensive and complex justifications, these are often automatically perceived as legitimate. This is even the case if the justification is incorrect. In other words, if it looks properly substantiated, it should automatically be correct, i.e., induced acceptance (Loi & Spielkamp, 2021; Sunstein & Gaffe, 2024). Third, AI taking autonomous action often leads to diffused accountability. In which developers, programmers, and users of AI systems only partially participate in the whole AI system lifecycle: in its design, programming, or usage (i.e., the problem of many hands; e.g., Bovens, 1998). They often possess only limited knowledge of the system or its results. As a result, leading to a dilution of the human responsibility behind AI action and, as a result, diffused accountability. These problems are not easily resolved within existing human-held accountability structures. Proving the urge for clarity in the right accountability mechanisms.

The previous challenges also portray a critical insight: AI should not merely be seen as an auxiliary component created to improve human decision making and agency. Instead, it can be seen as a separate independent actor capable of taking its own actions, actions that are subsequently being interpreted and overseen by human actors. This relationship can thus be visualized as two actors, one (human actor) overseeing the actions of the other (AI actor). A perspective that gains even more importance since AI is becoming increasingly more complex in the distant future, taking on tasks not limited to simple routine tasks. Therefore, this thesis set out to clarify the complex notion of accountability in the context of autonomous AI systems. To be specific, by critically analyzing whether accountability could be redefined in such a way that AI can also be held independently accountable.

Research into the possibility that accountability can be held by an AI actor has particular relevance to the accounting domain. As mentioned, accounting has traditionally played a central role in supporting accountability mechanisms within organizations, especially through standardized and formalized practices designed to enforce control and transparency. However, this dominant role is being questioned within critical accounting literature. This literature finds that human actors are inherently social beings, their identity shaped by social constructions

they are embedded in, e.g., by the social norms, languages, and relationships (e.g., Messner, 2009; Roberts, 2009). Humans are subconsciously exposed to this, leading to a part of their actions being opaque to themselves. As a result, they are incapable of providing full explanations and justifications. However, since that is often demanded from standardized and formalized accountability mechanisms within organizations, which are often supported by accounting systems, it leads to adverse reactions, such as self-centered or unethical behavior. These critiques have challenged the suitability of traditional accounting mechanisms for ensuring alignment with organizational and ethical goals. Instead, this leads to advice for softer, more informal forms of accountability, by embracing the deficiencies of human actors¹, and consequently decreasing the central role of accounting within accountability mechanisms and therefore organizations. On the contrary, if accountability is to be redefined in a way that AI systems themselves can also be held accountable, this may provide an opportunity to reassess the role of accounting within accountability mechanisms. AI possibly does not possess the relational weaknesses of human actors, namely. As a result, this could shift the position of accounting from the sidelines, no longer unfitting, back towards the heart of organizational accountability again.

Thesis question

As such, while HitL literature presents the human actor as uniquely capable of holding accountability, for instance, due to the capacity to provide transparency, it is the accounting literature that provides a conflicting portrayal. To be specific, it challenges the idea that human actors can provide a comprehensive account of their actions, and as a result, missing transparency. So, if the stated fundamental assumptions in HitL literature illustrate human accountability as perfect, why does previous accounting literature paint a rather different picture, seeing humans as limited in capacity? In addition, the argument of HitL also relies on the assumption that since AI is being programmed, it cannot be the originator of its own actions and cannot be held morally responsible. However, this thesis questions these assumptions, since they are rather narrow and one-sided. Instead, to gain a more nuanced perspective, this research introduces the philosophy of action. This philosophy explores concepts such as intentionality, consciousness, and free will in depth to determine, in a comprehensively informed way, if these assumptions in the HITL literature are still valid. Based on these considerations, the following research question is formulated:

¹ This form is commonly recognized as relational accountability, which will be discussed in Chapter 3.

Are the requirements for holding accountability exclusively human when viewed through the lens of the philosophy of action, and if not, what are the implications for the role of accounting in structuring organizational accountability mechanisms?

An answer to this thesis question provides theoretical relevance to the complex notion of accountability within organizations involving AI systems. Therefore, also offering practical relevance, this thesis can examine whether AI could be a viable bearer of accountability, thereby potentially bypassing these persistent issues with HitL mechanisms.

Outline of chapters

The structure of the remainder of this thesis is outlined below, along with the key claims advanced in each chapter: Chapter 2 illustrates the fragmented literature on AI ethics, philosophy of action, and accounting, revealing the lack of integration. It claims that only with a comprehensive synthesis of these different fields of literature can one truly assess whether AI can be held to account and what this conclusion could mean for accounting. This gap in the literature explains the basis and the motivation for this thesis. Chapter 3 involves the theoretical framework, defining the scope of key concepts in this thesis. Based on this delimitation, Chapter 4 delves into the different elements of accountability to assess if the argument of human uniqueness truly holds up. Therefore, this chapter assesses the assumptions within the HitL literature that assert the human uniqueness of accountability. This thesis claims that these are found to be ambiguous and grounded in narrow, one sided reasoning, but still finds certain prerequisites needed for holding accountability, such as intentionality, free will, and consciousness. Furthermore, this thesis finds that since the human actor is traditionally central to accountability, accountability practices are also formed and suited to human nature. As a result, this thesis sets out in this chapter a dominant trend in accounting literature to advocate for reducing the focus on accounting as a traditional practice for structuring and enforcing accountability, since they are assumed to be ill-suited. Instead, social practices are promoted as the solution. This thesis highlights how AI-held accountability could challenge this trend in the literature.

Chapter 5 draws on the philosophy of action to explore whether the truly filtered out prerequisites for accountability from Chapter 4 could be met by AI under certain defined conditions. These different philosophies are connected in the synthesis in Chapter 6, proposing a multi-layered framework that distinguishes between varying levels of interpretation with

regard to the concepts of intentionality, consciousness, and free will. Based on this framework, this thesis claims in the conclusion of Chapter 6 that AI can hold a specific form of organizational accountability, even better than human actors, named instrumental accountability. This thesis has done so by maintaining the classical structure of accountability, showcasing that it redefines the practical application of accountability, not the abstract notion of accountability. Consequently, it is claimed that an increase in AI usage within organizations could result in the resurgence of accounting as a primary practice for structuring and enforcing accountability. At last, chapter 7 involves the appendices, including the references.

Methodology

Instead of using an empirical approach, this thesis follows a purely theoretical approach. It allows for an in depth, comprehensive, and rigorous analysis of the current literature on a specific topic. Furthermore, this approach allows for critical comparison, reflection, and the development of a more nuanced understanding of the respective concepts. This thesis utilized key databases such as ACM Digital Library, IEEE Xplore, and Springer, via WorldCat and Google Scholar, using concepts such as ‘AI Accountability’ and ‘Human-in-the-loop²’. These databases enabled the collection of literature to identify dominant schools of thought within relevant literature fields such as accounting literature, HitL literature, and philosophy of action.

Rather than conducting a purely focused comparative literature study based on the full breadth of these fields of literature, this thesis is positioned as a study that critically evaluates the found conflict in relevant studies within these fields. Through this conflict, this thesis seeks to determine what the true essential prerequisites are for holding actors³ accountable within organizations. However, these abstract prerequisites found in this thesis, namely intentionality, consciousness, and free will, are still far from concrete, even subject to multiple (subjective) interpretations. As such, this recognition motivates the deliberate foray into the philosophy of action. This is the field of literature that is concerned with what these still abstract prerequisites entail, aiming to provide them with concrete meaning. As a result, it truly helps assess whether actors outside of human beings, like AI systems, can possess these as well.

² This includes all literature that provides arguments for the need for human oversight of AI action. Even literature that does not mention ‘Human in the loop’ directly is included in this classification. This is motivated by the limitation in the literature that directly mentions HitL as a concept. Nevertheless, the main focus must still be on the capacity of AI systems or similar advanced computer systems.

³ An actor can be both Artificial Intelligence (AI) and human in the terminology of this thesis.

2. Literature review

The current literature on redefining organizational accountability around AI related processes and its impact on the organizational role of accounting is only extensive when considered in fragmented forms across separate research fields. For instance, the majority of relevant literature connected to AI focuses on purely redefining accountability in AI systems, without its connection to accounting. This field of literature often assumes that AI cannot be held accountable, and typically focuses on redefining accountability to what constitutes the responsible human use of AI systems (e.g., Floridi et al., 2018; Raja & Zhou, 2023; Joseph et al., 2024).

Only a smaller subset⁴ of this literature actually engages with the fundamental question of whether organizational accountability can be redefined in a way that AI could hold it. This literature typically explores this question by focusing on one of the key elements of accountability: namely, answerability (e.g., Doshi-Velez et al., 2017; Miller, 2018), responsibility (e.g., Bartneck et al., 2020; Lechterman, 2022), or sanctionability (e.g., Stahl, 2006; Sparrow, 2007). These conclusions are mostly conservative about the abilities of AI to meet these elements, drawing only philosophical assumptions derived from thinkers like Kant (Bartneck et al., 2020) and Searle (Moor, 2006). This literature finds certain prerequisites⁵ unique to human actors, required to ensure responsibility.

Conversely, literature that arrives at more permissive conclusions about the abilities of AI often avoids a thorough engagement with these prerequisites. Instead, arguing that even human agents may fail to fully meet these prerequisites (e.g., Sullins, 2006⁶), or arguing that responsibility can be assigned even in the absence of such features, through pre-programmed, designed, or functionally embedded behaviors⁷ (e.g., Allen et al, 2000; Floridi & Sanders, 2004; Gips, 2011; Behdadi & Munthe, 2020). In both arguments, it tends to weaken the necessity of these prerequisites, instead of thoroughly engaging with whether AI can meet them. As a result, it leaves it unclear whether earlier conservative conclusions are justified, or

⁴ This is often seen as AI ethics.

⁵ The internal states of having consciousness, intentionality, and free will.

⁶ Robots may not have it, but we may not have it either, so I am reluctant to place it as a necessary condition for moral agency. (Sullins, 2006, p. 27)

⁷ This means it is believed that ethics can be implemented functionally, through programmed rules and constraints with, as a result, behavior that conforms to ethical standards without requiring internal moral understanding.

whether accountability itself might be redefined in a way that makes room for AI as a responsible agent.

Also, next to the literature that focuses on redefining accountability in AI systems, there is a completely separate field that engages with the implications of AI processes on accounting practices. This field explores the direct effect of changes in the accounting practice, i.e., how AI automates accounting practices instead of redefining the abstract organizational role of accounting (e.g., Leitner-Hanetseder et al., 2021; Boritz & Stratopoulos, 2023; Odonkor et al., 2024). Other literature within this field aims its research on redefining accountability to what the responsible human use of AI systems is. This mainly aligns with the field of AI ethics, but this time for accounting professionals (e.g., Schweitzer, 2024; Cao & Zhang, 2025). While both research directions are valuable, they do not explore what the impact of potentially holding AI accountable on its own could mean for the institutional role of accounting within organizations. That is, its role as a system for structuring and enforcing organizational accountability.

These previous fields of literature provide relevant information but remain fragmented and one-sided, as a result, lacking integration and critical comparison, and as a result, missing a clear conclusion based on the different perspectives. Therefore, this thesis challenges the deficiencies in current contributions of AI ethics, with respect to both the conservative but also the permissive argument, by highlighting and critically analyzing their often very selective philosophical view and mentioning them. This critical analysis is conducted by bridging the different perspectives of AI ethics together and thoroughly reflecting on these perspectives via philosophy of action (e.g., Anscombe, Dennett, Frankfurt). In this way, simultaneously, this thesis also complements the conclusions of the existing contributions to the permissive field of AI ethics, by creating a better structured and more comprehensive argument in favor of AI.

Consequently, via this critical reflection, a clear contribution can be made to the accounting literature, to both accounting literature contributions connected to accountability and AI. Namely, this thesis challenges the prevailing trend in accounting literature toward advocating relational accountability within organizations. Instead, it reconsiders the value of instrumental accountability in the light of increasing AI organizational dominance. In contrast, it mainly complements the field of accounting in connection to AI, by providing a whole new view, not focused on how accounting directly changes due to AI automation, or how we should responsibly use AI in accounting. Instead, it explores how accounting as a practice itself

becomes more central for structuring and reinforcing accountability within AI-integrated organizations.

3. Theoretical framework

The theoretical framework defines the scope of key concepts within this thesis. While this chapter discusses the most significant ones requiring delimitation, it is by no means exhaustive in the key concepts being discussed.

Artificial Intelligence

Concepts such as algorithmic systems, artificial intelligence (AI), and automated decision making (ADM) systems are conceptually interconnected in the literature but are not always clearly distinguished. These concepts are, for instance, being used interchangeably, such as in HitL literature (e.g., Loi & Spielkamp, 2021; Raja & Zhou, 2023) or in the literature on algorithm aversion (e.g., Araujo et al., 2020; Hou & Jung, 2021; Cheng & Chouldechova, 2023). As an example, Araujo et al. (2020) state that ADMs can be both algorithmic systems and artificial intelligence, based on how the concept is framed⁸. While correct, the relationship between the concepts is better understood as one being a subset of the other, seeing algorithmic systems as functional ADM systems. In other words, ADMs act as an umbrella term for different specialized technologies, with AI systems as a subset of it. This is, for instance, acknowledged by the EU AI Act (European Commission, 2021). While multiple forms of selective definitions of algorithmic systems exist, for instance, Dietvorst et al. (2014) only include systems used for forecasting, the specific definition of Sunstein & Gaffe (2024) is much more comprehensive. They refer to the concept as an algorithm that (1) takes a set of inputs, (2) conducts some set of computations, and (3) generates an output that may consist of predicted outcomes, probability assessments, synthesized analysis, summary information, or recommendations.

⁸ An automated decision-maker can be conceptualized as an algorithm, a recommender system, or, simply an “artificial intelligence” depending on how it is framed and presented to the user of the system or subject of the decision. (Araujo et al., 2020, p. 613)

AI systems, however, as a subset of algorithmic systems, according to the EU AI Act, are referred to as:

A machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. (European Commission, 2021)

One can dissect multiple key elements of AI from this definition, illustrating the intelligence of this algorithmic system that sets it apart from other algorithmic systems. First, while discussions around AI in the organizational context are often framed as humans augmented with AI (e.g., Boritz & Stratopoulos, 2023), AI is designed to take autonomous action, particularly without direct human help or intervention (e.g., Poole & Mackworth, 2010). This means even within HitL frameworks, this autonomy is managed, rather than eliminated, through oversight or ex post evaluation. This ability to make certain autonomous moves that affect real-world outcomes will be referred to in this thesis as technical autonomy. Second, this action is based on explicit or implicit objectives. This means it acts within a programmed scope, based on explicit decision rules, or by using pre-determined models to accomplish certain tasks and targets (Kaplan & Haenlein, 2019). Third, what adds to the intelligence layer in addition to the autonomous part, this action exhibits adaptiveness by using different development technologies, mainly including machine and deep learning. These are abilities that enable systems to develop themselves further via analyzing, adapting, and improving performance based on collected data without actual programming (Goodfellow et al., 2016). Finally, this adaptive autonomous action has an actual influence on physical or virtual environments, an influence comparable to that of a human, meaning it can be seen as an artificial agent (e.g., Russell & Norvig, 2010; Poole & Mackworth, 2010; Bartneck et al., 2020).

Based on the previous characteristics, AI can be seen as an artificial agent that interprets external data and takes adaptive (technical) autonomous action on its own within a defined scope to achieve specific goals. As a result, AI is the most advanced subset of algorithmic systems, capable of mimicking intelligent human behavior (Russell & Norvig, 2010; Bartneck et al., 2020). For instance:

1. An AI system independently creates financial dashboards by selecting and calculating relevant KPIs, accompanied by a brief summary.
2. A human overseer consequently verifies the results and takes the AI-informed financial recommendation to the board.

As a result, processes involving at least a significant AI system involvement have certain actions taken autonomously without direct human input. In other words, for every step in a process, either the human acts or the AI does. This example also illustrates the binary distinction in autonomy. In part 1, the technical autonomy rests with the AI system. In part 2, the technical autonomy rests with the human actor. Although part 2 is AI-informed, the final decision power rests in human hands. To conclude, this thesis distinguishes two forms of organizational action: AI action and human action (potentially augmented by AI information). As a result, within this thesis, a distinction is made between human actors and AI actors⁹. This delimitation is made to overcome the complexity of humans integrated with technology, delving into the philosophy of technology, a distinct research domain on its own.

Accountability

Next to AI, understanding accountability in its narrow definition for this thesis is also of considerable importance, considering the wide and vague usage of the concept in literature. To illustrate this, two examples can be formulated, taking inspiration from Lechterman (2022):

1. Someone is being held accountable by his neighbor for commonsense noise pollution.
2. Someone is being held accountable by his manager at work for his professional actions.

Two examples illustrating that accountability in its basic meaning can be applied in multiple settings. Accountability, as a concept, is often used as an abstract, often ambiguous umbrella term (Mulgan, 2000; Bovens, 2010). To be able to completely grasp the meaning of accountability, it is necessary to return to its historical and also philosophical roots. Taking the notion from Hartmann (2023), in both ancient biblical theology (e.g., May, 1946), as in Aristotle's ethics (e.g., Moss, 2014), the Latin concept of 'logos' captures the essence¹⁰ of

⁹ From this point onward, this thesis refers to artificial intelligence systems as 'AI' or 'AI actors,' and to human beings as 'humans' or 'human actors'.

¹⁰ As such, literature interprets 'logos' the same way as accountability (e.g., Barlev, 2006).

accountability. Biblical theology describes ‘logos’ as the divine word or the moral imperative to answer your actions before a higher power (God). Aristoteles describes it as the capacity to reason, to understand, and to provide an explanatory account of one’s actions. This illustrates that the root lies within the capacity to formulate an account instead of the obligation to do so. As such, it also describes the terminological combination of ‘account’ and ‘ability’ into accountability. This previous understanding sets the foundations for later literature on accountability. For instance, a commonly used basic meaning of accountability in literature entails:

Accountability refers to the implicit or explicit expectation that one may be called on to justify one's beliefs, feelings, and actions to others. (Lerner & Tetlock, 1990, p. 255)

While this description explains still the bare minimum, literature often introduces another element: one may expect to face positive or negative consequences set by others, for the quality of one’s justification and actions based on the evaluation of these actions according to some normative standards (e.g. Stenning, 1995; Mulgan, 2000; Bovens, 2007). These sentences suggest three different elements: justification, a counterpart, and the expectation of consequences. Using these elements, one can dissect the difference between examples 1 and 2. The first example illustrates a counterpart without a position of power, not being able to enforce this accountability, only capable of assigning praise and blame by informal moral standards (Lechterman, 2022). Instead, in the second example, the counterpart has a higher power (in this case, a hierarchical power by contract, while this could also be a regulatory power. This power (such as the government) has the ability to enforce its request to take an account from one another and evaluate one's account by formal standards, and apply consequences.

Furthermore, Bovens (2007) not only acknowledges the presence of formal consequences as a crucial part of accountability, but he also dissects the key element of justification into two separate parts: the expectation to provide an explanation, but also the expectation to provide a justification. While explanation refers to the sharing of information on how one’s conclusion is reached (which is descriptive in nature), justification, rather, is the argumentation and legitimization of its actions and choices (which is normative or moral in nature). Together, they can be seen as the account.

Last, a concept often being used interchangeably with accountability is responsibility. However where being held accountable for is mainly focused on the previous description, responsibility is mainly described as the condition that one is taking intentional actions and being recognized as the originator of those actions; in conclusion, one is responsible for it. Responsibility can be seen as a sub-element of accountability; however, some papers recognize accountability as an element of responsibility (e.g., Lechterman, 2022). The reason for this is that they are interconnected: responsibility implies also being liable for actions and thus capable of facing consequences. As a result, responsibility automatically implies the elements of accountability.

Nevertheless, the dominant view in accountability literature is seeing responsibility as the last missing element in accountability. While Chapter 3 further explores the concept of responsibility in depth, this thesis draws a clear distinction between moral and legal responsibility (liability) for the scope of the research due to its pragmatic implications. By law, a person or organization is responsible where an AI actor does not have this legal capacity; consequently, while this thesis could write a potential desire to, it can never be currently legally held accountable. However, a better understanding of the philosophical question if human accountability is inherently unique in a moral sense could inform legislators in the development of new laws, illustrating the relation between moral and legal responsibility.

Combining these key elements in a three-element conceptual framework, used by e.g., Raja & Zhou (2023) and Loi & Spielkamp (2021), accountability refers to the expectation that one can be called to:

1. Explain and justify their actions (answerability)
2. for his own intentional actions (responsibility)
3. to face consequences set by a hierarchical power (sanctionability)

If accountability meets all three key elements, it is referred to as strict external accountability, deriving its name from the presence of a hierarchical higher power holding one accountable.

Next to the fully realized form, literature also identifies different abstract conceptualizations of accountability based on the absence of the third key element (e.g., Mulgan, 2000; Bovens, 2010; Lechterman, 2022):

1. In the absence of formal consequences: *accountability as responsiveness*
2. In the absence of justification and thus also formal consequences: *accountability as (symbolic) control*
3. In the absence of oversight with higher power and thus also formal consequences: *accountability as dialogue, forensic accountability*
4. In the absence of the counterpart and thus also oversight and formal consequences, *accountability as a virtue, internal accountability*

While all these abstract conceptualizations of accountability differ in terminology, they share a common core. Without the third key element, i.e., without a higher power that is capable of setting consequences, multiple risks could occur. Accountability could become subjective and vague, or even overcomplicated due to the absence of a strict hierarchical power taking account from one (Mulgan, 2000). As a result, accountability can lose its practical function to hold one formally and effectively accountable. This thesis makes a distinction between strict external accountability, meeting all three key elements of the framework in a formal sense, and other, more informal or symbolic forms.

This thesis is focused on providing a better understanding of accountability around AI in accounting related processes, as AI increasingly takes on autonomous roles in accounting-related processes. As a result, this thesis is mainly focused on the understanding of accountability within the organizational context. Research on the understanding and right implementation of accountability (mechanisms) in an organizational context is traditionally dominated by a focus on strict external accountability and criticism of the other forms (e.g., Jones, 1992; Mulgan, 2000). Advocates of this form argue that ethical and compliant behavior must be incentivized or enforced through strict monitoring, standards, and sanctions. Traditional organizational accountability thus offers a very structured form based on external accountability, by being grounded in institutional roles, formal reporting lines, and hierarchical mechanisms of answerability, responsibility, and sanctioning. In organizational practice, this is seen as an instrumental/hierarchical form of a concrete accountability mechanism.

In contrast, more progressive views in accounting literature rather argue for a less strict external accountability form. These views are mainly focused on conceptualizations of accountability beyond external accountability, in the absence of the sanctionability element of accountability (e.g., Sinclair, 1995; Roberts, 2009; Vosselman, 2016). In organizational

practice, this is seen as a relational/socializing, i.e., a collective form of concrete accountability mechanism. Advocates of this form argue that compliant and ethical behavior mainly forms through informal spaces where actors hold each other accountable through dialogue, relationships, and shared responsibilities since humans are social beings. A form where multiple abstract interpretations of accountability apply: as a virtue, or as a dialogue, for instance. These interpretations have a shared core, not meeting the third key element, and are more informal and symbolic in nature. This is developed as an answer to the argued downsides of strict instrumental accountability, such as the self-centered culture and moral disengagement (e.g. Roberts, 1991).

Based on the previous two forms, Roberts (2009), for instance, proposes the idea of a hybrid form, calling it intelligent accountability. Which seeks to combine the strengths of both instrumental and relational mechanisms. Similarly, Bovens (2007) argues that we should include accountability as a strict control mechanism (instrumental) and promote accountability as a virtue (relational) within organizations. Bovens suggests, as a result, that both are necessary and complementary. Intelligent accountability acknowledges the importance of social structures for ethical behavior while also emphasizing the necessity of formal, strict mechanisms.

Since both traditionally strict external accountability mechanisms and progressive relational accountability mechanisms are being utilized in organizational contexts, it is ultimately up to the organization to decide which form of accountability is preferred and deemed as more effective. Both mechanism types are relevant to the accounting domain, given that accountability mechanisms can be utilized to design concrete and suitable structures for accountability in accounting. However, it is important to note that accounting itself is also a practice of this in specific instrumental accountability. In everyday operations, accounting involves tracking actions and justifying outcomes, also actively implementing these instrumental accountability structures toward hierarchical oversight. As a result, gaining a deeper understanding of accountability around AI is useful in two key ways: it informs the design of appropriate accountability mechanisms with regard to AI in accounting, and it could redefine the role of accounting as a central instrument in upholding instrumental accountability.

In conclusion, the three element conceptual framework of accountability forms represents the common base in literature and the central basis of this thesis. It does not require the presence of a formal third key element of accountability, since it acknowledges both instrumental as relational accountability forms within the organizational context for this thesis. However, the presence of the two other elements is required: an actor must be responsible for their actions and capable of providing an account of them. These elements can be used to develop a concrete understanding of the complex notion of accountability, in order to evaluate the presumed uniqueness of human accountability in comparison to AI's incapacity to be held accountable in HitL literature.

4. Assessment of key elements

HitL literature mainly elaborates on the correct and responsible design and implementation of human accountability mechanisms surrounding the usage of AI, rather than considering the possibility of AI being held responsibly accountable. This is based on optimistic assumptions about human actors capable of meeting the key elements of accountability, and rather pessimistic assumptions about AI actors to meet these key elements.

By contrast, this chapter highlights shortcomings with both these assumptions of human actors and AI actors. First, the lack of in-depth argumentation underpinning the assumptions of AI actors in this HitL literature, with, as a result, only a select few sources engaging the topic at a deeper level. Second, the conflict in views between HitL literature on human actors in comparison to influential accounting literature focused on accountability. As a result, this shows that HitL literature does not provide enough clarity in argumentation to demonstrate that AI cannot be held accountable universally. Therefore, this chapter dissects these different assumptions and critically reflects on them per core element of accountability: answerability, responsibility, and sanctionability, to conclude per element whether the argument for human uniqueness truly still holds up.

Answerability

According to HitL literature, accountability mechanisms and organizational structures should account for limitations in answerability in AI. AI is seen as a black box, as being opaque and having low transparency, explainability, and intelligibility capacities (e.g., Bartneck et al., 2020; Lechterman, 2022; Raja & Zhou, 2023; Joseph et al., 2024; Schweitzer, 2024). In other words, these papers assume that AI is not capable of explaining and justifying its actions. Instead, arguing for human oversight, i.e., a human in the loop, to fill the missing capacity of answerability. This is clearly illustrated by the following citation:

Moreover, the opacity of AI algorithms raises concerns about transparency and accountability, making it challenging for stakeholders to understand and trust AI-driven decisions. [...] Human-AI collaboration is essential for achieving effective information governance, and defining clear roles and responsibilities is critical for establishing a framework where human judgment complements AI capabilities. (Joseph et al., 2024, p. 112)

Upon deeper analysis, most HitL literature, however, does not elaborate further on the assumption that AI does not have this capacity, nor what they actually mean when suggesting that AI is a black box. To be able to answer these questions, one must understand what is concretely desired when discussing the requirement of answerability to be held accountable as the first key element. In the Oxford Handbook of AI Governance, for instance, Lechterman (2022) simply states that AI processes are just opaque to human observers, without elaborating on this statement. In addition, they state that even when AI actors are more transparent in the information behind decision making, this is often still unintelligible to humans. While not touching further in depth on the assumption behind it, it touches on an important notion: answerability is not about transparency, it is about explainability and intelligibility.

Transparency is often used as an interchangeable concept with explainability or intelligibility (e.g., Dietvorst et al., 2014; Kaplan & Haenlein, 2019; Kroll, 2020; Hou & Jung, 2021). The significant difference between transparency versus explainability and, even more, intelligibility is, however, clearly illustrated in literature: “Suppose software containing millions of lines of code is made transparent, what would be the benefit of this?” (Bartneck et al., 2020, p. 36). Explainability, for instance, is not to be equated with transparency. While transparency does provide a counter to the black box principle of AI (providing better insights into the programming codes), it does not automatically mean this code could be easily and correctly

interpreted (Kroll et al., 2016; Bartneck et al., 2020). Especially the more complex the decision becomes, the more unreadable the code becomes, losing its usability (Kroll et al., 2016; Lechterman, 2022). As a result, it does not automatically explain the decision making process of an AI actor. This, in conclusion, explains the distinction between transparency and explainability.

Explainable AI (XAI) research aims to address the challenges of transparency and explainability, which is often described as “opening the black box”. The main objective of this research involves the production of AI models that are better understood via extensive explanation processes (Ali et al., 2023; Reddy & Kumar, 2023). These explanations are often very technical, repeatable, and standardized since these are mainly aimed at providing crucial information to developers and data scientists for quality control and trust in legal compliance (Dwivedi et al., 2022). This includes, for instance, the technique of “feature rankings”, which helps explain which input variables influenced a decision the most. This standardized output of these feature rankings can be repeatedly logged, compared, and audited across different cases. As a result, it mainly illustrates explanations that are not as easily interpretable to every user.

Literature also illustrates an often seen as an even more significant concept for the answerability element of accountability, being distinct from explainability. This concept is named intelligibility. Explainability refers to the actor’s ability to provide the information and logic used to arrive at their decisions (Doshi-Velez et al., 2017; Miller, 2018). Intelligibility, however, refers to the perception of these explanations to different groups of actors; it can be for instance, a strong explanation in technical terms for programmers, but can still be an unclear explanation (and thus unintelligible) for direct users (e.g. Bartneck et al., 2020; Lechterman, 2022). Intelligibility refers thus specifically to the capacity to provide both an explanation and justification for one’s actions in a way that is personally intelligible to the one holding the actor to account. However, when summarizing previous notions, the literature stays relatively superficial with regard to the case of answerability. The HitL literature often argues incorrectly for transparency as the main goal instead of intelligibility. Also, even if mentioning intelligibility, it rarely engages with the deeper meaning of this concept.

So the question remains: what is unique about human answerability concerning intelligibility, one that AI is not capable of according to the HitL literature? Select HitL literature does actually delve into depth into this notion. An intelligible explanation and justification entails one that involves the factors and their weights that influenced the specific decision, and in addition, by expanding on the counterfactual (Doshi-Velez et al., 2017; Miller, 2018). This counterfactual explains what would be required to reach a different outcome by giving an answer to the question: “What if this factor were something else?”.

Also, Doshi-Velez et al. (2017) argue that human actors are capable of reflecting on their decisions *ex post*. This means that even if a factor has not been included yet in the decision making and explanation of a human actor, this actor can revisit its decision and reflect if this new or previous unknown factor that would have influenced the decision making. This connects to the notion of the counterfactual: “If I had known this factor, I probably would have decided otherwise, or not”. These are the precise abilities of a human actor that can be appealed to by the recipient of the explanation/justification¹¹ (the one calling the actor to account), making it intelligible for the recipient. This means that different recipients can ask different unknown or unexpected factors to provide a counterfactual explanation for, thus showing the significance of human actors capable of providing flexible explanations *ex post*.

It is thus the capability of flexibility that makes an account intelligible to the recipient holding one to account. What is particularly interesting is that this does not imply that these intelligible accounts are often rationally correct. Citing Doshi-Velez et al. (2017, p. 5): “[...] Humans are notoriously inaccurate when providing post hoc rationales for decisions”. Human accounts are often not fully accurate reflections of the original reasoning process. Humans might even rationalize their decisions after the action, using reasons in other ways or that were not consciously part of the original decision process. This is motivated by the desire to produce tailored accounts, based on what the explaining human actor believes is most relevant to the recipient, often leading to biased accounts. These biases do not occur due to dishonesty, but because human reasoning is adapted to social communication and efficiency (Miller, 2018). To be specific, a human actor is shaped and influenced by their relational environment, which makes them prone to produce a tailored, yet often biased and imperfect account. At the same time, this relational sensitivity enables them to read the recipient’s background knowledge,

¹¹ From this point onward, ‘explanation’ and ‘justification’ will be referred to as ‘an account’ collectively.

expectations, and values, to be able to produce a socially intelligible account. The human deficiency in producing a perfectly accurate account enables thus a human actor to provide a truly intelligible account. These findings by Doshi-Velez et al. and Miller reflect a broader understanding of human rationality, aligning with Herbert Simon's influential theory of bounded rationality. This theory holds the idea that human decision making is limited due to cognitive and environmental constraints, sensitive to context, and as a result, aims for only satisfactory outcomes in line with the context rather than optimal outcomes (e.g., Simon, 1990). Therefore, it can be seen as a form of bounded rationality. The human actors, namely, aim to satisfy the demand for a uniquely tailored account, even in the absence of full accuracy and the objective truth.

Most HitL literature thus stays relatively superficial on the understanding of why AI actors are unable to meet the capacity of answerability, often suggesting transparency issues as the main problem. Only a select few authors touch on the notion of intelligibility, let alone in depth. This further underscores the relevance of drawing on accounting literature, which offers a further in depth exploration of the mechanisms underlying human answerability. Mainly substantiated by authors with a progressive view. A point of view that does not see human actors as rational individual beings, but rather as relational beings embedded within a socially constructed collective: the social norms, languages, and relationships (e.g., Shearer, 2002; Messner, 2009; Roberts, 2009). These authors build their argumentation on the psychology of Judith Butler, who in turn draws on the work of Emmanuel Levinas. It is especially the philosophy of Butler that provides the foundation for progressive accounting literature and is simultaneously perfectly in line with the HitL literature regarding intelligibility. Butler's notion can be perfectly understood via the following citation:

[...] if it is precisely by virtue of one's relations to others that one is opaque to oneself, and if those relations to others are the venue for one's ethical responsibility, then it may well follow that it is precisely by virtue of the subject's opacity to itself that it incurs and sustains some of its most important ethical bonds. (Butler, 2005, p. 20)

Butler (2005) argues in her book 'Giving an Account of Oneself' that based on this socially constructed collective, one's identity is shaped. It is thus shaped by cultural and other interpersonal forces one did not choose directly ourselves, but instead by one's environment. These forces one is exposed to subconsciously lead to a part of one's actions being socially

conditioned, which is opaque to oneself. Butler, as a result, challenges the idea that human actors can give a full, rational, and comprehensive account of one's actions. One often takes action based on subconscious social norms, languages, and relationships. Thus, forcing a full account from someone, a demand of full self-identity is seen as a certain ethical violence according to Butler, one human actor cannot meet.

The very terms by which we give an account, by which we make ourselves intelligible to ourselves and to others, are not of our making. They are social in character [...] (Butler, 2005, p. 21)

This citation highlights that not being able to fully explain oneself is not seen as a moral flaw. Instead, it is rather the starting point for the possibility of ethical relations and also being intelligible in accounts. Human actors are social beings, fully living in their social constructs. A notion in line with Doshi-Velez et al. (2017) and Miller (2018), representing the select few HitL authors delving into depth into intelligibility.

Based on this notion of ethical violence, Messner (2009) states that it makes organizational accountability demands to human actors intrinsically problematic, especially if they require neutral or comprehensive justification. He sees it as paradoxical, the higher the accountability demands for comprehensive accounts, the less transparent human actors become. Instead of fostering openness, it leads to human actors withdrawing or strategically shaping their accounts. As a result, undermining the goal of answerability and thus accountability.

[...] they establish social norms, a domain of unfreedom and substitutability within which our 'singular' stories are told. (Messner, 2009, p. 924)

The citation illustrates this paradoxical problem, mainly arising from instrumental accountability mechanisms. Instrumental accountability mechanisms are often standardized rules and metrics, which do not consider the limitations in human answerability to account for the opaque self (based on one's subconscious social construction). Instead, these mechanisms are standardized to allow for substitutability between employees (to treat actors like interchangeable parts). As a result, there is no freedom (thus a domain of unfreedom) to speak freely for oneself; instead, one's "singular" self is fitted through these organizational standards. These instrumental accountability mechanisms thus do not account for the true requirements of human intelligibility. Therefore, influential authors within accounting literature, such as Shearer (e.g., 2002), Messner (e.g., 2009), and Roberts (e.g., 2009) advocate for replacing

accounting as a central practice for structuring and reinforcing accountability within, towards practices beyond these formal metrics. These involve mutual recognition, face-to-face dialogues, and other social obligations. As a result, this dominant view in accounting literature thus stresses the need to decenter accounting as a central mechanism within accountability structures, decreasing organizational relevancy.

What is particularly noteworthy is that instrumental accountability mechanisms are structured in the same way as XAI: both are standardized, repeatable, and designed to enable substitutability. Where instrumental mechanisms reduce the human actor to a compliant substitutional unit, XAI explanations reduce the account to a set of technically legible and compliant outputs. This structural similarity confirms that XAI is indeed unable to inherently meet the deeper relational requirements of intelligibility. As a result, not meeting relational accounting mechanisms within organizations.

In conclusion, the human ability to provide intelligible accounts in contrast to AI clearly supports the HitL claim of human uniqueness over AI actors. This is particularly relevant when organizations endorse relational accountability and prefer mechanisms aligned with it. Nevertheless, while the HitL literature argues that AI actors are a poor fit for holding organizational accountability, it must also be acknowledged that human actors frequently struggle to meet these same accountability demands in organizations. A tension that is especially found in organizations endorsing and preferring instrumental accountability. This is never explicitly mentioned in the HitL literature; instead always sees human accountability as the unquestioned solution.

This tension provides, as a result, already clarity for the complex notion of accountability around AI actors, namely for the first key element of accountability: answerability. If organizations believe in the benefits of instrumental accountability and decide on this form: demanding full transparency based on standardized rules, XAI might even be able to meet these demands, which humans are rather ill-suited for. As a result, XAI can be held accountable according to the instrumental answerability demands. If organizations believe in and endorse relational accountability or hybrid forms¹²: they desire intelligible, socially grounded accounts

¹² From this point onward, this thesis refers to ‘relational’ or ‘hybrid’ accountability mechanisms as both relational forms.

that are flexible ex post and tailored around the recipient. As a result, something that XAI and other systems cannot currently meet. However, this relational mechanism is designed to accommodate the limitations of human nature: the vulnerability and deficiencies in their accounts. As such, it may be unreasonable to expect AI to conform to these standards at all.

In conclusion, the distinction between instrumental and relational mechanisms helps clarify how organizations might approach the question of answerability around AI processes. Organizations with a more traditional¹³ mindset may find that explainable AI (XAI) is an even better fit for the element of answerability than human actors, who are often unsuitable to such rigid demands. In contrast, organizations that currently embrace relational forms of accountability mechanisms may find that AI is rather poorly aligned with these mechanisms if held answerable. Consequently, if these organizations still seek to hold the AI actors accountable, they may need to reconsider a shift back toward more standardized and formalized instrumental answerability mechanisms.

Responsibility

While the answerability provides an interesting duality in possibilities for AI holding accountability itself, responsibility is also needed to fulfill the full spectrum of accountability. The HitL literature states, as it does with the answerability element, that AI can never be inherently responsible for the outcome, and therefore can never be held accountable. A commonly unquestioned notion in the HitL literature is:

In AI ethics guidelines, it is normally assumed that only human persons can be accountable, (current) AIs cannot. (Loi & Spielkamp, 2021, p. 10)

So instead, the HitL literature suggests that accountability mechanisms and organizational structures should be designed in such a way as to ensure the human responsible use of AI (e.g., Floridi et al., 2018; Raja & Zhou, 2023; Joseph et al., 2024).

¹³ Those who endorse instrumental accountability, favoring transparency, standardization, and formal justification.

Similar to answerability, one must understand what is concretely desired when discussing the requirement of responsibility to be held accountable as a second key element. The HitL literature defines and refers to different concepts when discussing responsibility, to be specific: liability, autonomy, free will, and agency. For instance, the Handbook of Ethics in Robotics and AI (Bartneck et al., 2020) defines responsibility as the attribution of praise or blame to an agent for an outcome, by asking the question inherently: “Who can be blamed for the outcome?”. The Oxford Handbook of AI Governance (Lechterman, 2022) defines the notion of responsibility as forensic, by reflecting ex post to provide praise or blame. This praise or blame can be seen as a duty to answer for one’s actions (e.g., Dignum, 2022). While this defines responsibility in its basic form according to HitL literature, mainly three different forms are highlighted in the literature:

1. Causal responsibility (e.g., Kroll, 2020): descriptive in nature, simply referring to the factual causation of what led to the outcome. X caused Y, without any moral judgement attached.
2. Legal responsibility, i.e., responsibility as a duty (Kroll, 2020), i.e., liability (e.g., Bartneck et al., 2020): by law and/or must be proven in court, referring to the legal form.
3. Moral responsibility (e.g., Coeckelbergh, 2020; Bartneck et al., 2020), i.e., responsibility as a fault (Kroll, 2020): complex and philosophical in definition.

Causal and legal responsibility are both 2 binary in nature, meaning something or someone has caused an outcome or not, referring to causal responsibility, or one is proven by law or in court to be legally responsible and thus liable for an outcome or not. Both are defined by objective and clear criteria. Moral responsibility, however, as it involves questions of morality and ethics, is much more subjective and philosophical in its nature. The differences between these forms can be further illustrated via the following citation:

Traditionally in AI and robotics, the term autonomy refers to an AI system’s or robot’s ability to operate without human intervention. In this section, however, we focus on the ethical principle of autonomy. (Bartneck et al., 2020, p. 30)

An AI actor has the ability to take technical autonomous action, which has actual influence on physical or virtual environments. In other words, it causes things: having causal responsibility. The question of who is actually morally responsible and consequently also potentially liable (legally responsible) for these actions is not necessarily the same as the actor who is causally responsible. For instance, an AI actor may autonomously flag and block a high value bank transfer of a company because it detects it as a suspicious transaction. In this case, while AI is causally responsible for the financial damages that occurred due to this block, it is still undetermined who is morally or legally responsible. Only technical autonomy thus equals at most causal responsibility. What can also be stated is that legislation around legal responsibility or liability is often informed by the understanding of moral responsibility. This concludes that HitL literature mainly refers to the moral form of responsibility as the fundamental element of accountability (e.g., Coeckelbergh, 2020; Bartneck et al., 2020).

The HitL literature connects multiple different concepts to moral responsibility, often, however, lacking sufficient argumentation. Namely, an actor can only bear moral responsibility if one has, in addition to technical autonomy also moral autonomy¹⁴ (e.g., Moor, 2006; Coeckelbergh, 2010; Floridi et al., 2018; Lechterman, 2022), free will (e.g., Moor, 2006; Bartneck et al., 2020), intentional action, i.e., having an intentional state or intentionality (e.g., Moor, 2006; Johnson, 2006; Lechterman, 2022), consciousness (e.g., Moor, 2006; Stahl, 2006), or emotions like empathy, i.e., having an emotional state (e.g., Stahl, 2006; Coeckelbergh, 2020). These concepts are often used as reasoning why it is unquestioned if AI can be held morally responsible, without actual, extensive argumentation. For instance:

Much like nonhuman animals, AI cannot be responsible because it does not have the relevant intentional states. (Lechterman, 2022, p. 166)

Lechterman does not further expand on this assumption, not even what intentional states are, leaving the reader wondering why this is automatically assumed. Instead, HitL often sees AI actors as limited to purely technical autonomy, with the purpose to support and enhance human decision making, i.e., optimized human agency, so called smart agency, where AI actors are only seen as organizational artifacts (Johnson, 2006; Floridi et al., 2018; Dignum, 2022). Citing Johnson: “[...] computer systems do not have the intention to act arising from their freedom.”

¹⁴ The terms ‘autonomy’ and ‘agency’ are often used interchangeably: literature refers to both ‘technical agency’ and ‘moral agency’, as to ‘technical autonomy’ and ‘moral autonomy’, often within the same contexts.

(Johnson, 2006, p. 203). The HitL literature is mainly focused on optimizing the responsibility around the lifecycle use of these artifacts.

Nevertheless, even though the HitL literature often lacks structured and in-depth argumentation to support the assumption that AI cannot be held morally responsible, it does present some engagement with the underlying concepts. To be specific, the concepts of free will, consciousness, and the presence of intentional and emotional states. Starting with free will, Bartneck et al. (2020) expand on their idea of free will, via their reasoning on the philosophy of Kant:

As autonomous beings, Kant holds, we are obliged to rationally consider our moral standards. Simply following the law is not good enough. It follows that an artificial intelligent system must have autonomy in the Kantian sense to be able to act in an ethical way. (Bartneck et al., 2020, p. 32).

Bartneck et al. (2020) suggest that because AI is preprogrammed in a certain way, it is bounded in its decision making and thus unable to consider its actions freely and choose for itself. As a result, it thus lacks free will in the Kantian sense. Bartneck et al. take an Autonomous Weapon System (AWS) as an example of an AI actor. If the AWS system is preprogrammed and consequently acts and decides based on these bounds created by humans, this implies it cannot be responsible for itself for killing the wrong person. Instead, they assume that humans are free of these programmed bounds, have free will, and should be held morally responsible.

Where Bartneck et al. (2020) mainly touch on free will, the other HitL literature (e.g., Moor, 2006) explores in depth the concept of intentionality (i.e., the intentional state) by referring to the philosophy of John Searle. To expand on this, according to Searle (1980), an intentional state is an actor's own beliefs, desires, and intentions. He suggests that this state can only be formed by a conscious mind since an actor must experience and understand what it is doing for its actions to have their own beliefs, desires, and intentions. Just as the biological phenomenon of photosynthesis requires the specific biology of plants, consciousness and intentionality are seen as biological phenomena arising from the specific biology of a human brain. Searle (1980) argues that actors that are "programmed", thus never replicating physically the true biological structures of a brain, such as the synapses, will never have true intentionality or consciousness.

Instead, when asked whether an artificial actor would recreate the true structure of a brain, would it possess consciousness and true intentionality?:

Assuming it is possible to produce artificially a machine with a nervous system, neurons with axons and dendrites, and all the rest of it, sufficiently like ours, again the answer to the question seems to be obviously, yes. If you can exactly duplicate the causes, you could duplicate the effects. And indeed it might be possible to produce consciousness, intentionality [...] (Searle, 1980, p. 422)

So his answer to the question is positive, suggesting that a conscious mind with intentionality can be created as long as it has the same biological structure. It provides some philosophical base, suggesting that the fundamental underpinning of human uniqueness finds itself in the biological nature of the human brain, resulting in certain metaphysical powers. A philosophy that is often called biological naturalism. This philosophy touches on the core assumption underlying most HitL literature: all suggesting in diverse ways that the programmed and artificial nature (without a brain) of AI creates a bounded capacity.

Related to this abstract consciousness is another reason as an argument that AI cannot hold moral responsibility, which is the incapacity to feel emotions such as compassion, fear, care, or forms of empathy, that activate the ability to be moved by moral concerns (Stahl, 2006). Often recognized as the strong emotion view, borrowed from the philosophy of Martha Nussbaum (Coeckelbergh, 2020). Based on these emotions, like empathy, one can make intentional moral choices that ‘feel good’, a mental state only an actor with consciousness can have (Stahl, 2006; Coeckelbergh, 2020). Without emotional mental states, decision making would just fall into the hands of “psychopathic rule-following robots”, that is, not moved by moral concerns themselves (Coeckelbergh, 2020). As a result, it thus requires consciousness to create an emotional state, to have emotions like empathy within decision making.

“We typically regard humans as having consciousness, intentionality, and free will.” (Moor, 2006, p. 20). A list of traditional individual conditions, all assumed AI actors do not have (Stahl, 2006). While not touched extensively in depth by HitL literature, it does refer to philosophers such as Kant or Searle to better substantiate their argument for their assumptions about human uniqueness. One can conclude that this literature refers to several fundamental concepts as prerequisites for moral responsibility: such as free will, consciousness, and intentionality. However, they also point all to a common conclusion: due to its programmed

and digital nature, AI lacks a consciousness to have emotions such as empathy, it lacks intentions and lacks a free will, which makes it unable to be held morally responsible.

Just as with answerability, accounting literature provides valuable insights into the complex notion of accountability with regard to the second key element of accountability: responsibility. While the HitL literature often draws from the philosophical traditions that define responsibility in terms of prerequisites rooted in the individual self, accounting literature takes a different approach. Instead, accounting literature highlights how responsibility is especially shaped through a relational dimension: social relations, institutional norms, and organizational roles, again referring to Butler and Levinas (e.g., Shearer, 2002; Messner, 2009; Roberts, 2009). What is thus particularly interesting is that, although they do not see human beings as programmed actors, they also simultaneously reject the idea that responsibility emerges from their own autonomy as mentioned in the HitL literature. One can delve deeper into Butler's and Levinas' understanding of the uniqueness of human responsibility, quoting Butler:

Responsibility is not a matter of cultivating a will, but of making use of an unwilled susceptibility as a resource for becoming responsive to the Other. (Butler, 2005, p. 91)

There is no moral responsibility in the traditional Kantian autonomous sense that is based on an individual's free will and self-reflection. Instead, Levinas argues that responsibility arises as an affective response based on the moral demands of the Other¹⁵. Human actors are vulnerable to these demands, precisely because they are social beings, part of social constructions shaped by shared norms and expectations. As a result, acting on this demand without reflection. Responsibility thus emerges pre-reflective and unwilled, not by choice but because one is affected by the Other. Human actors are conditioned, formed, and moved by external structures, norms, and relations they did not choose themselves. According to this view, having a consciousness, an emotional state, and free will is not sufficient to meet true moral responsibility. Instead, an actor needs to be part of a social structure, vulnerable to the demands of others, to be seen as truly morally driven. For instance, one is moved to care, even before one can choose to do so. Empathy is thus in this case not a voluntary emotion, but rather already an automatically affective response to the moral demand of the Other.

¹⁵ Levinas and Butler refer to 'the Other' as any person within the social construction that can make a moral demand upon a human actor. It is thus not a specific individual but rather a placeholder for all others we are morally vulnerable to.

As a result, based on this philosophy, accounting literature (e.g., Shearer, 2002; Messner, 2009; Roberts, 2009) suggests that organizations should create and cultivate environments that enable and encourage relational accountability. Only in this way, by stimulating a shared culture with relational structures via dialogue and values, human actors can bear and foster moral responsibility. It is believed by this philosophy that if this is crowded out, which is often the case in pure individualistic instrumental accountability mechanisms, human actors become unable to express moral responsibility.

In conclusion, the definition of responsibility goes beyond the distinction in the definition of answerability. While answerability reflects a clear distinction: relational mechanisms are better suited to human actors, and instrumental mechanisms are better suited to AI actors, the element of responsibility is more driven by beliefs. To be specific, traditional views are based on the assumption that compliant and ethical behavior can only be assured via incentivization and enforcement through monitoring, standards, and sanctions. This is classified as the strict external form of accountability, which still relies on the assumptions that actors must be seen as the originators of their actions, to be held responsible and therefore as a result accountable. To be specific, it requires that an actor possess the individual prerequisites: intentionality, consciousness, and free will. This aligns, in conclusion, with the arguments of HitL literature, which argues that without the individual prerequisites, no moral responsibility can be attributed, and as a result, an actor cannot be held accountable.

In contrast, organizations that endorse relational forms of accountability mechanisms, organizations believe that compliant and ethical behavior can only be assured via social constructions shaped by shared norms and expectations. In this case, the individual prerequisites alone are not sufficient. Instead, responsibility emerges only when the actor is taking part in these social relations, via dialogue, for instance. This contrast reveals that moral responsibility is not a fixed quality, but depends on the assumptions that organizations hold about responsibility and relations. If responsibility is understood in individualistic terms, only the individual prerequisites must be attainable for AI. However, if responsibility is emerging from participation in social contexts, then AI would need to meet much more demanding criteria.

As a result, the threshold for fulfilling the requirements of responsibility is lower within instrumental accountability than within relational accountability. Still, AI would need to fulfill the individual prerequisites of moral responsibility to be able to classify actions as its own.

While HitL slightly touches on this in depth via the philosophy of Searle, suggesting that it is currently impossible due to the biological naturalism argument, it offers only a narrow, one-sided view. Therefore, chapter 5 will explore a broader philosophical landscape by including the philosophy of action, providing nuance by investigating whether AI can meet these individual prerequisites.

Before moving to the final discussion regarding these prerequisites, the last component of accountability, named sanctionability, must be considered. While this is not required for relational accountability mechanisms, it is especially of great significance in the context of instrumental accountability. Since the instrumental accountability mechanisms are a form of strict external accountability, they require the element of sanctionability for external enforcement. As a result, AI actors should be capable of being sanctioned and held instrumentally accountable.

Sanctionability

The last element of accountability involves sanctionability. Sanctionability entails the capacity to impose consequences on actions based on the quality of one's justification and actions, which are evaluated against some normative standards (e.g., Stenning, 1995; Mulgan, 2000; Bovens, 2007). These consequences are mainly formal, for instance, punishments or penalties, and function as part of instrumental accountability mechanisms, used to enforce compliance with organizational goals (e.g., Mulgan, 2000). However, when referring to sanctionability in the broader sense, also informal consequences are included, such as praise or blame based on shared moral standards (Lechterman, 2022). These moral consequences are typical of relational accountability, but lack the enforceability, and thus are not considered formal within accountability mechanisms. To have an effect, these consequences must have moral meaning to the actor. The actor must be capable of feeling emotions like fear and happiness based on these consequences. This emotional impact, in turn, can foster an internal learning mechanism through the feedback and clarity these consequences provide (e.g., Bovens, 2007; Roberts, 2009).

As a result, the element of sanctionability with regard to its capacity can be broken down into three different parts. The capacity to: set consequences, have emotional impact, and evoke an internal learning mechanism. Where literature on the first two accountability elements is relatively extensive, considerably less is written about the sanctionability of AI. Nevertheless,

this does not imply that AI cannot be sanctioned according to the HitL literature, but it is rather divided over the possibilities. This chapter delves into the three different parts of sanctionability to analyze if AI can be sanctioned. With regard to the first part, the capacity to set consequences, HitL literature is relatively unanimous. One can state that, in a purely practical sense, it is possible to sanction, e.g., punish, an AI actor. This can be understood as a form of pragmatic sanctionability. For instance, citing Lechterman (2022):

[...] AI can be accountable for wrongdoing because it can be sanctioned by modifying or deleting it.
(Lechterman, 2022, p. 166)

However, with regard to the second part of sanctionability, the capacity to experience moral and emotional impact from the consequence(s), the HitL literature is less optimistic about, acknowledging the limitations of its impact (e.g., Stahl, 2006; Sparrow, 2007; Bartneck et al., 2020). For instance, citing Bartneck et al.: “[...] what does it mean to 'punish' a system that cannot feel or suffer?” Bartneck et al. (2020, p. 42). Sparrow (2007) argues that without the capacity to feel emotions such as sympathy, grief, or remorse in a way that is morally compelling, there is no real moral sanctioning. We can apply pragmatic consequences, but for a sanction to count as a real moral consequence, it must have moral meaning for the one being held to account. This, in turn, requires consciousness: the actor must be able to feel emotions, like fear. This can be understood as a form of moral sanctioning. Citing Sparrow:

In order for our treatment of the machine to count as punishment, it must be capable of suffering in ways that might motivate the same set of responses that we have as a matter of course to human beings.
(Sparrow, 2007, p. 72)

This emotional impact should have, in turn, the capacity to foster an internal learning mechanism, through the feedback and clarity these consequences provide. While it can be said that since AI does not have a consciousness, consequences do not evoke an emotional impact, nor foster an internal learning mechanism. This can be understood as a form of functional sanctioning. However, as depicted in the theoretical framework, AI already exhibits adaptiveness in its nature, via machine and deep learning, which enables systems to develop themselves further via analyzing, adapting, and improving their performance and capabilities (e.g., Goodfellow et al., 2016). This suggests AI has the capacity to reflect and learn from its actions and feedback, independent of the capability to receive sanctions.

This emotional impact should have, in turn, the capacity to foster an internal learning mechanism, through the feedback and clarity these consequences provide. While it can be said that since AI does not have a consciousness, consequences do not evoke an emotional impact, nor foster an internal learning mechanism. This can be understood as a form of functional sanctioning. However, as depicted in the theoretical framework, AI already exhibits adaptiveness in its nature, via machine and deep learning, which enables systems to develop themselves further via analyzing, adapting, and improving their performance and capabilities (e.g., Goodfellow et al., 2016). This suggests AI has the capacity to reflect and learn from its actions and feedback, independent of the capability to receive sanctions.

In conclusion, since sanctionability is a key element in instrumental accountability mechanisms, it should meet the requirements of the respective element. While AI can be sanctioned pragmatically and also functionally, it does not evoke a learning response based on the moral and emotional impact, because it lacks the capacity for moral sanctioning. On the contrary, AI actors still exhibit learning mechanisms based on feedback. As a result, it becomes debatable if AI is thus truly sanctionable without moral sanctionability. However, the prerequisites for moral sanctionability are similar prerequisites for the element of responsibility within instrumental accountability. This means it can be concluded that the capacity to be held accountable comes down to still only the three prerequisites.

This underscores the importance of exploring the landscape of the philosophy of action in depth in the next chapter to investigate whether AI can truly meet these individual prerequisites. As a result, concluding whether AI can be correctly held independently accountable to instrumental accountability norms, thus providing clarity on the complex issue of accountability in AI usage.

5. Philosophy of action

This chapter sets out different influential authors within the philosophy of action. To be specific, the names of Anscombe, Dennett, and Frankfurt each offer different perspectives on the fundamental prerequisites for holding true instrumental accountability: intentionality, consciousness, and free will. This means that while notions from these different philosophers are used, it does not provide a comprehensive utilization of their full theories, which are too complex to elaborate on completely. Instead, selected interpretations of the prerequisites are used to form a structured understanding per philosophy. By exploring these concepts per author, this thesis proposes a result: a multi layered framework that distinguishes different levels of interpretation for what is considered true instrumental accountability. Based on these different interpretation levels, the framework serves to examine whether AI actors can meet the prerequisites: intentionality, consciousness, and free will, which have traditionally been reserved for purely human agents.

Anscombe

“What distinguishes actions which are intentional from those which are not?” (Anscombe, 1957a, p. 321). A citation characterizing the main focus of Anscombe’s philosophy on the concept of intentionality. Although Anscombe does not address the other two prerequisites necessary for being accountable, her focus on understanding intentionality remains influential and foundational. As such, this forms a valuable theoretical entry point before assessing and understanding the more complex (but also complete) frameworks of Dennett and Frankfurt. Anscombe (1957a, b) sets three sorts of explanations for actions apart. Mental causes, intentions, and motives. A mental cause, as the concept describes, refers to the direct cause for the action: a feeling, perception, emotion, impulse, or other similar trigger that produced the reaction to act. Anscombe provides an example:

[...] ‘Why did you knock the cup off the table?’ answered by ‘I thought I saw a face at the window and it made me jump’. (Anscombe, 1957a, p. 321)

This explains the immediate cause of the action, the trigger; it does not represent an explanation for action based on true goal direction or awareness of action based on deliberate reasoning and judgement. Often due to a product of your feelings, perceptions, and/or impulses.

On the contrary, intentions are explanations to Anscombe, where the deliberative answer can be given to the why?-question. It refers to something done on purpose, under a description. One can ask: Why are you doing that? It thus shows there must be awareness of acting, being rational, and goal-directed. However, Anscombe acknowledges, this can be minimal/implicit and very simple. The goal can be the action itself, but one must be aware of the reason. It may be preceded by a mental cause or without, as long as there is an intention. For instance, one is running to make sure to catch the bus on time as an intention, which can be preceded by the image of seeing the bus already stopping as a mental cause; however is still valid in the absence of the mental cause. This, as a result, explains the following citation by Anscombe:

It is not in all cases that 'I did so and so in order to...' can be backed up by 'I felt a desire that...'
(Anscombe, 1957a, p. 324)

At last, Anscombe distinguishes motives from mental causes and intentions:

Motives may explain actions to us; but that is not to say that they 'determine', in the sense of causing, actions. (Anscombe, 1957a, p. 327)

Anscombe thus argues that the concept of intentionality is not based on authenticity, autonomy, or free will, but rather on the ability to provide a clear answer and reasoning for the action itself, an account. As such, she actually suggests that intentionality, while treated as a prerequisite for responsibility, is in fact rooted in the capacity for answerability. To be specific, it refers to the ability to provide a self-grounded account of one's actions, not per se requiring that it is intelligible to others.

In conclusion, what should be acknowledged is the lack of engagement with concepts such as consciousness and free will, which an actor must possess to truly bear moral responsibility and become sanctionable. However, Anscombe's philosophy clearly sets out different actions that can and cannot be understood as intentional. She rethinks intentional action not as an internal mental state but rather as a particular functional form of reasoning, a clear conceptual shift. As a result, she illustrates and enables a more accessible understanding of this shift, one that Dennett subsequently extends and complicates in his own framework even beyond intentionality.

Dennett

Compared to Anscombe, Dennett provided a more comprehensive philosophical account of intentionality, consciousness, and free will. Dennett starts with consciousness in his earlier work (e.g., Dennett, 1969), which stands in direct contrast with Searle's philosophy. Dennett (e.g., 1969, 1991) proposes a rigorously different conceptual framework of consciousness, seeing it not as one phenomenon, feature, or aspect, but rather as a combination of several ones working together (Dennett, 1969). There is no little man in our head that runs our mind, nor one central theater where everything happens: metaphors illustrating his clear rejection of a central abstract metaphysical consciousness.

Instead, consciousness is depicted as a decentralized system with clear functional components, whereas in his initial model, *Content and Consciousness* (1969), he mainly dissects it into two functional components. In this initial model, he describes the functional component of awareness (the processing of 'simple' information) and the functional component of introspective reportability (ability to communicate this awareness) (Dennett, 1969). Deliverances (i.e., via verbalizing) are thus, as a result, proof that there was a certain awareness at a certain time, being communicated via the introspective reportability. Citing Dennett:

The human capacity for making introspective reports is seen as a mode of access to the content of awareness, and in virtue of the invulnerability to error examined in the last chapter, its deliverances are seen as reliable – indeed conclusive – evidence of the content of awareness. (Dennett, 1969, p. 119)

Not only does Dennett dissect this metaphysical understanding of consciousness into functional components, but he also alters his focus from personal to sub-personal level: in order to understand bodily functionalities working together. The consciousness can be seen as constantly giving possible interpretations to new awarenesses via the nervous system and evaluating this meaning via different evaluation methods, such as the human perception or memory system. Based on this sub-personal level, Dennett expands his consciousness framework into the multiple drafts model:

According to the Multiple Drafts model, all varieties of perception— indeed, all varieties of thought or mental activity—are accomplished in the brain by parallel, multitrack processes of interpretation and elaboration of sensory inputs. Information entering the nervous system is under continuous "editorial revision". (Dennett, 1991, p. 111)

A draft is a potential meaning to an awareness, where most drafts become rejected in the process and only some persist and give meaning, becoming part of the memory, the emotional state, or becoming deliverances, for instance, via verbal behavior (such as speech). These drafts that persist are the conscious contents. Consciousness is thus not one center, but rather a combination of sub-personal processes working together. This is what makes the consciousness; it is like the center of gravity: it does not exist in science, it is theorists' fiction. Instead, it is just a conceptual abstraction to explain the motions in the world, just as in our head, so it can be called the center of narrative gravity.

Next to consciousness, in *Content and Consciousness* (1969), Dennett explored the concept of intentions and intentionality, seeing it as something separate from consciousness. For instance, while a common understanding is that animals do not possess (the same level of) consciousness as humans do, Dennett still suggests that animals have intentionality:

Anything that suffices to demonstrate this must itself be suspect, for we do describe animal behaviour in intentional terms quite successfully. (Dennett, 1969, p.132).

Intentionality is the aboutness of mental states, in other words, their goal direction towards things (Dennett, 1988). It does not require awareness and introspective reportability, but rather to be predictable via the intentional stance. A conceptual framework which he elaborates on in his later work, *The Intentional Stance* (1987), and furthermore optimized in *The Intentional Stance in Theory and Practice* (1988), and later on in *Intentional Systems Theory* (2009), understanding a fundamental concept again from a functional perspective, rather than an inner metaphysical capacity. As Dennett argues, intentionality can exist without consciousness, but intentionality provides a foundation for the existence of consciousness. As a result, this relates to free will and responsibility. Citing Dennett:

I do not intend to present a 'solution' to the problems of responsibility and free will here, but certainly a first step in any such solution must be finding the crucial difference between intentional and unintentional actions. (Dennett, 1969, p. 172)

The intentional stance explains an actor's behavior and actions as if it is the result of rational choices grounded in an actor's beliefs and desires, even if it needs pretending. Dennett argues it is of lesser importance if this is inherently correct, but rather if the strategy has predictive power. To be specific, one can make consistent and correct predictions about what the actor is going to do next. Which, in turn, is based on the (possibly incorrect) assumption that it has beliefs about the world, has certain desires or goals, and acts rationally to fulfill its desires based on these beliefs. If these predictions are consistent and correct, even if these assumptions are incorrect, the actor still has functional (derived) intentionality. Dennett's pragmatic and functionalist view aligns closely with the principle of positive economics, created by Milton Friedman (1953). Both argue that the predictive power of a model or stance is more important than the realism of the underlying assumptions. It complements the argument of Dennett, that it is pragmatically justified through its predictive capabilities rather than the underlying truth of the assumptions.

One can even state that a thermostat has derived (as-if) intentionality: an intention to control the temperature to follow a specific temperature goal. On the contrary, we can also aim the intentional stance strategy at a human actor, where we do not need to use an as-if form of intentionality, but where true, original intentionality can be stated (Dennett, 2009). Particularly noteworthy is that there is no magical cut-off point between the two forms of intentionality, on the scale of actors between a simple, mere thermostat and a complex human actor. Instead, it is seen as a continuum with "[...] no theoretically motivated threshold distinguishing the 'literal' from the 'metaphorical or merely 'as if' cases" (Dennett, 2009, p. 8). Instead, as actors become more complex, their intentionality becomes deeper and more literal. This complexity is measured in functional competence, to be precise, citing Dennett:

The robot poker player that bluffs its makers seems to be guided by internal states that function just as a human poker player's intentions do, and if that is not original intentionality, it is hard to say why not. (Dennett, 2009, p. 9)

In summary, intentionality is a continuum based on as if (derived) towards original intentionality. The more complex and competent a system becomes, the more original intention it goes towards, except we can never know; we can only use predictive tests. Similar to the concept of consciousness, both are depicted as functional mechanisms, without any metaphysical parts. Where intentionality comes first in a functional sense: it's about being able

to predict behavior through goal-directedness, and assumed beliefs, which could even if programmed still be present and seen as authentic intentionality.

With these philosophical standpoints, Dennett rejects the traditionally accepted Cartesian dualist interpretation of human consciousness: the view that human consciousness is something non-physical substance separate from the brain. A clear dualism between the mind and the body, where dualists see the mind as a centralized inner place where ‘it all comes together’, according to Dennett, which he calls the Cartesian theater. Instead, he sees it as all kinds of sub-personal functionalities working together:

There is no single, definitive "stream of consciousness," because there is no central Headquarters, no Cartesian Theater where "it all comes together" for the perusal of a Central Meaner. (Dennett, 1991, p. 257)

What is particularly noteworthy is that this ought not to be of relevance, given the fact that not only Dennett but also Searle (1980) is rejecting this Cartesian dualistic philosophy. Intentionality and consciousness are a biological phenomenon that emerges from the neurophysiological, i.e., biochemistry of the brain. So, Searle argues, they cannot be seen as separate:

For example, I reject dualism, so he says I believe in the soul. I think it is a plain fact of nature that mental phenomena are caused by neurophysiological phenomena [...] (Searle, 1980, p. 454)

As a result, both accusing each other. Searle argues that seeing consciousness and intentionality emerge from something else than a biological, neuro-physical, structured brain assumes duality. Dennett argues that Searle portrays some sort of metaphysical ‘magic’ non-replicable mind, making it again dualistic according to the Cartesian philosophy (the classic metaphysical mind-physical body).

While consciousness and intentionality are relatively unambiguous in Dennett’s philosophy, the concept of free will remains much more opaque, lacking a clearly defined conceptualization. Therefore, an interpretation of his free will in this thesis is based on multiple statements found in his literature. A concept of much significance, as Dennett highlights:

If it turns out that there is no such sort of free will, the very idea of responsibility will lose its foundation. (Dennett, 1984, p. 167).

Dennett argues that it is often assumed that we must reject the possibility that we live in a deterministic world to have free will. Otherwise, in such a worldview, every event is assumed to be fully caused by prior events. As a result, it is often concluded that an individual could not have acted in another way. Therefore, no true free will and moral responsibility would be possible. Instead, Dennett argues that even within a deterministic world, there is enough ‘elbow room’ to be morally responsible and make a difference in the world, since completely undetermined decision power is not needed (Dennett, 1984). The understanding of this elbow room can be interpreted via the following two citations:

This is just a particularly clear case of what we all always want: lots of elbow room. We want a margin for error; we want to keep our options open, so that our chances of maintaining control over our operations, come what may, are enhanced. (Dennett, 1984, p. 69)

A controls B if and only if the relation between A and B is such that A can drive B into whichever of B’s normal range of states A wants B to be in. (Dennett, 1984, p. 57)

This elbow room thus centers on the idea of having both control and, in addition, competence, to be able to steer outcomes within the constraints of a bounded world. If an AI actor possesses technical autonomy, it demonstrates control over its actions. Also, if it exhibits complex intentionality, such as that of the robot poker player, with structured beliefs, desires, anticipation, and learning capabilities (the intentional state), it can be said to have competence as well. However, Dennett argues that these two conditions alone, control and competence, are not sufficient for real freedom in the form of elbow room. He states:

[...] the capacity for conscious recognition of motivation is apparently a necessary condition of real freedom. (Dennett, 1984, p. 41)

In other words, without functional consciousness (the capacity of self-awareness), an actor can never truly exercise freedom. Instead, it may have intentionality, but it lacks the self-awareness required to revise and reflect on its behavior. This is especially the case within the constraints of a possibly deterministic world shaped by uncertainty and cognitive limitations, as Dennett calls it. It is this capacity, the ability to recognize one’s own intentional state, that

allows one to take elbow room in this deterministic world and create partial freedom for oneself.

From this standpoint, Dennett also challenges the philosophy of Kant, often referred to by HitL literature (e.g., Bartneck et al., 2020). Kant argues that full rationality is essential in decision making to be able to evaluate one's choices and have free will. Instead, Dennett embraces a more bounded view. Rather than undermining moral responsibility, cognitive limitations, and uncertainty become the motivation for what makes deliberated reflection meaningful and responsibility possible. We are not expected to be fully rational, since it would make reflection and responsibility hollow if an actor were perfectly rational already.

In conclusion, Dennett redefines the concepts of intentionality, consciousness, and free will not as metaphysical absolutes like Searle assumes: the binary consideration if it has natural consciousness and intentionality (having a little man in our head) or not. Instead, it is about pragmatic functionality. What matters is whether an actor can act with autonomy (control), functional intentionality (competence), and functional awareness (consciousness) in a way that supports moral evaluation and thus responsibility. Dennett thus argues that the question is not whether an agent meets the biological naturalist requirement, but whether it has the functionalities and behaves as if it has.

Frankfurt

As Dennett already acknowledged and introduced, our ability to form complex intentions, often via beliefs and desires about our own desires, depicts a higher form of intentionality, i.e., second-order mental states (Dennett, 1991). While Dennett did not elaborate further in depth on these second-order states, Frankfurt provided a comprehensive understanding of them, depicting the uniqueness of human action via its free will:

It is my view that one essential difference between persons and other creatures is to be found in the structure of a person's will. (Frankfurt, 1971, p. 6).

It seems to be peculiarly characteristic of humans, however, that they are able to form what I shall call 'second-order desires' or 'desires of the second order. (Frankfurt, 1971, p. 6).

Frankfurt thus touches on a key concept within the moral responsibility element of accountability. He distinguishes mainly:

- First-order desires: i.e., desires to do or have something.
- Second-order desires: i.e., desires about first-order desires

This division depicts what makes someone a moral person is not just that they have desires and act, but that they are aware of their first-order desires, can reflect on these, and have desires based on their first-order desires. One can choose which to endorse in the end, where one identifies with, which encapsulates the ability of personhood. That one is able to identify with one of the desires, which one chooses to endorse (Frankfurt, 1971).

In the case that someone is aware of their own first-order desires, but chooses not to identify with this, but rather chooses to endorse a second-order desire about this, and actually acts on it, it is called a second-order volition. Vice versa, one can also choose not to act on second-order desires if one actively chooses with his will to follow his first-order desire. No volition, but still aligns with one's personhood. In both cases, since one chooses actively, it is seen as free will: the capacity to form personhood based on multiple orders of desires, identify with one of them, and act on them.

In contrast, Frankfurt (1971) also recognizes actors who are missing (parts of) this capacity. Unwilling actors are able to form a personhood but are prone to their first-order desires and unable to actually act on them. An example of this is actors prone to an addiction: they have a second-order desire to reject the urge of the addiction and identify with this, but still fall for it since they cannot control their free will. In other words, it is the awareness that you have first-order and second-order desires (identification with them), and you are conscious of which one you want. If you choose the second order, then a volition, or not. But if you want the will to be second-order volition, but still choose automatically for first-order without wanting it, then you do not have complete free will.

This also shows the importance of control over intentions in Frankfurt's theory. So this means if you endorse your first or second-order desire, but still are not able to put it into action, you are acting unwilling, a product of your impulses. Having an intention does not automatically mean you have control over your actions; only when one is able to follow one's personhood

does one have free will. Consequently, one has difficulties accounting for their actions if it does not match their beliefs or endorsed desires, meaning if we act unwillingly. It is a lack of self-identification with our actions (Frankfurt, 1976). However, what is important to recognize is that Frankfurt argues that it is the personhood and self-identification, via reflection, that already means an actor is capable of moral responsibility, even if he cannot act on it (yet). An actor is capable of forming second-order volitions. Citing Frankfurt:

It is not true that a person is morally responsible for what he has done only if his will was free when he did it. (Frankfurt, 1971, p. 18)

This means even if one acts unwillingly and as a result, does not identify with one's own actions, one still has moral responsibility. Frankfurt argues it is more within the conscious and aware reflection of one's desires and the inherent deep self-identification with one's reasons, desires, and beliefs that makes one truly moral, rather than the capacity to act on it. Frankfurt (1971) also distinguishes a third group, called wantons. This is a group of actors who are only following their first-order desires, without being aware of it, by forming second-order desires about it. While they can act with intentionality, they do not engage in moral reflection via forming second-order desires and self-identification.

As a result, Frankfurt introduces a deeper dimension required for moral responsibility in comparison to Dennett. While Dennett focuses on functional control, competence, and awareness of one's desires, including the ability to reflect on them in the light of uncertainty and cognitive limitations, Frankfurt considers this insufficient: reflection without deeper self-identification is merely 'wanton' behavior. Instead, Frankfurt argues, reflection is a rather much more complex mechanism, used for self-identification and forming personhood with one of the often conflicting desires an actor holds. The ability to truly endorse one desire over another and the intention to act on it is what constitutes free will and grounds moral responsibility.

To conclude this chapter, the forays into the philosophy of action play a critical role in this thesis, as they reveal the significant variation in how the key prerequisites for accountability are interpreted. These differences in interpretations not only occur significantly in contrast to the narrow assumptions found in HitL literature, by thinkers such as Kant and Searle, but also within the philosophy of action itself. Namely, there are even significant differences in

interpretations between Anscombe, Dennett, and Frankfurt. As such, it provides this thesis with a clear concluding argument, but the individual prerequisites are not clearly objective or universally defined. Instead, they are prone to interpretation, completely depending on the individual philosophical lens through which they are viewed. Without these forays into the philosophy of action, this thesis would be limited to purely the narrow views of mainly Kant and Searle.

6. Synthesis and conclusion

This thesis finds that there could be mainly two different forms of organizational accountability distinguished: traditional instrumental and progressive relational forms. When HitL literature discusses the incapacity of AI to be held accountable, they mainly refer to the missing of certain prerequisites that align with instrumental accountability. However, this thesis argues that these prerequisites have only been explored in a narrow and one-sided manner. To gain a deeper understanding of what it truly means to possess these prerequisites, multiple philosophers of action have been discussed to get a comprehensive understanding.

Synthesis of Philosophies

Based on the exploration of these three influential philosophers of action, this thesis proposes a multi-layered framework that distinguishes between varying levels of interpretation with regard to the concepts of intentionality, consciousness, and free will.

- Level 1: Functional instrumental accountability

The first level, based on the philosophy of Anscombe and Dennett's functionalist approach, introduces a layer of instrumental accountability requirements that are based on the internal functional state of the actor. An actor must not just merely be able to explain its decisions but functionally behave like it has a complex intentional state with consciousness and free will. It must function as if it possesses beliefs, desires, goals, and the capacity to truly revise them in response to feedback. Only if an actor is responsive to reason, via awareness of its own decisions, and learn from outcomes, can it be said to possess consciousness and free will. Accountability becomes, in this view, possible when the actor is capable of being responsive to reasons: aware of its own actions, sensitive to norms and feedback, and revise its behavior accordingly (via self-reflection and learning). This means it must be capable of error correction

and self-monitoring. Current AI models aspire to meet this level of functioning via reinforcement/deep learning systems (Goodfellow et al., 2016). Only with this functional behavior, an actor can be said to possess complex intentionality, consciousness, and free will via the functional interpretation of ‘level 1’.

- Level 2: Authentic instrumental accountability

Level 2 further builds forth onto the functional model of level 1: not only must an actor have self-awareness and self-reflection in its behavior, but Frankfurt suggests it must also exhibit second-order reflection. An actor must be capable of having desires about desires and truly identify with a specific desire. Only when an actor is capable of endorsing this, the actor has the capacity to bear authentic moral responsibility and, as a result, to hold accountability. While level 1 is relatively clear if they can be met by current AI developments, it becomes uncharted territory with regard to level 2: authentic instrumental accountability. Namely, it pushes the design of AI far beyond current architectures. The accountability question is, in this case, no longer about answerability and rationality (via awareness and reflection), but rather whether an AI can truly identify with its decisions and possess an authentic intentional stance. One can ask where the distinction lies between simulated identity and genuine evaluative personhood. As a result, it does not especially characterize a definitive frontier for current AI technology, but rather a grey zone: can an AI actor genuinely identify with its decisions, or is it merely simulated coherence?

- Level 3: Biological instrumental accountability

At the highest end, Searle’s biological naturalism sets an objective, clear boundary. To be specific, unless an actor possesses a biologically neuro-physiologically structured brain composed of neurons and synapses, no true consciousness and intentionality can emerge. As a result, an actor cannot genuinely be the originator of its actions, and as a result, not be held accountable. Where level 2 is becoming more ambiguous in contrast to level 1, it is becoming rather unambiguous again in level 3: setting a clear objective standard, an AI actor must be created with a true biological artificial brain to be held accountable.

This framework mainly illustrates that instrumental accountability within organizations is not a simple, singular option but rather a combination of multiple options. Also, these options are not simply shallow and purely preference based decisions but are instead ethically and practically loaded. As a result, if wanting to solve the problems of HitL structures within AI

processes via assigning instrumental accountability to AI actors, organizations are facing an organizational dilemma. Organizations must ask themselves what level they are willing to accept as morally and institutionally sufficient for instrumental accountability: one that can fully ensure that actors align with organizational and ethical goals, while still maintaining efficiency and effectiveness.

On the one hand, choosing a lower threshold (e.g., level 1) results in the need for less costly and complex AI models, but simultaneously risks the ethical integrity and legitimacy of the accountability mechanisms and, as a result, structures within the organization. For instance, if an AI actor meets the prerequisites without deeper self-identification, stakeholders may question whether accountability is truly upheld. If a system lacks true personhood, the system can still provide technically correct justifications and act functionally moral, but lack genuine commitment to ethical principles. It creates a hollow form of accountability, one that is no longer focused on substance but rather on shallow appearance. As a result, it could weaken the legitimacy of these structures in the eyes of the stakeholders of the organization.

On the other hand, choosing a higher threshold (i.e., level 2 or 3) results in the demand for significantly more advanced AI models, which are exponentially more costly to develop. While these models align more closely with deeper conceptions of responsibility and accountability, they may also come at the risk of ambiguity (i.e., level 2). It becomes unclear whether AI can genuinely endorse its own goals or merely simulate that endorsement. As a result, even though Level 2 aspires to a deeper ethical grounding, it also opens up unresolved philosophical and technical uncertainties. As a result, the framework of instrumental accountability is not a morally indifferent construct. The threshold an organization adopts reflects not only a practical consideration (organizations should weigh the efficiency of investing in these advanced AI actors against the benefits of avoiding problems accompanied by the reliance on HitL mechanisms) but also a moral posture. Namely, if organizations see accountability as a narrow compliance mechanism, or as a structural commitment to moral responsibility, accountability, and ethical coherence.

Conclusion

This thesis was conducted to provide some clarity on the complex situation of accountability around organizational AI processes. Specifically, it challenges the narrow and one sided assumptions about AI's capacity for holding accountability. As a result, it also challenges the view that accounting should be decentered as the primary practice for structuring organizational accountability in all modern organizations. In order to provide an answer to the research question, it requires to be broken down into several components, each reflecting a layer of this thesis research to accountability itself.

First, the dissection between the two forms of accountability mechanisms. Traditionally, organizations have relied on instrumental accountability mechanisms, which strictly encompass all three key elements of accountability. This thesis concludes that AI actors (i.e., XAI), by design, may be better capable than human actors of providing consistent, comprehensive, and standardized accounts of their actions. In fact, existing literature argues that human actors are instead often unfitting to meet these comprehensive and standardized demands. However, this thesis finds that AI still needs to meet the responsibility and sanctionability elements of accountability. These elements rely on fundamental individual prerequisites, including intentionality, consciousness, and free will. All three of these individual concepts are required to be truly morally responsible and sanctionable.

Literature also sets out a more progressive form of accountability mechanism, named relational accountability. This form is designed to accommodate the limitations of human nature: the vulnerability and deficiencies in their accounts, making sure that the mechanisms still steer behavior towards organizational and ethical goals, instead of adverse behavior. This thesis concludes that it stands the farthest from AI actors and should not be the goal of organizations to design AI around this form of accountability mechanism for three reasons. First, the core problem that relational accountability aims to solve, namely the risk of adverse behavior resulting from instrumental mechanisms, is assumed not to apply to AI actors as it does to human actors. Second, implementing AI held relational accountability would require the development of complex layers of social responsiveness, going beyond only meeting the previous individual prerequisites. Third, instrumental accountability already aligns naturally with the design of AI actors, making it a more practical fit as a result.

Referring back to the research question, are these prerequisites actually unique to humans if viewed through the lens of the philosophy of action? This thesis concludes that this is not a simple yes or no question. Instead, a multi-layered framework is developed that helps clarify under which conditions AI or other actors, past humans, could possess these prerequisites as well. Each level offers a different threshold for deciding what is required of an actor to be held meaningfully (instrumentally) accountable. As a result, this framework provides organizations with a conceptual tool to navigate the complex moral terrain of accountability. However, these different levels are not simply practical choices but are also significant normative commitments. They force organizations to morally consider what kind of moral responsibility is good enough to substitute human actors, which could often lead to organizational tension. This tension is characterized by a trade-off. Choosing a lower threshold will provide greater cost-efficiency in model development. However, this results in accountability being seen as a procedural justification method only, being superficial in moral responsibility. In contrast, if a higher threshold is chosen, it can preserve this moral depth, and as a result, also stakeholder trust in these mechanisms. However, this requires more complex systems, which potentially even exceed current technological limits and also result in significantly higher development costs. In conclusion, if AI actors are developed to be as equally complex as humans, thus meeting level 2 or level 3, they will be, by design, better suitable for holding instrumental accountability than human actors.

Therefore, what does this imply for accounting? This thesis concludes that once AI actors and other artificial actors become increasingly more central to organizational processes and decision making, this thesis argues that instrumental accountability should also become more central. This means it could be even the default again, as a practice for structuring and enforcing organizational accountability. This thesis concludes this shift back to traditional standards based on multiple concluding arguments that have been set out in the previous chapters. Namely, as modern organizations grow in AI usage, it results in increasing problems with oversight, such as the mentioned interpretability issues and the problem of many hands. However, since this thesis concludes that AI could be held to account, and if organizations seek this accountability as a solution to the oversight challenges, instrumental accountability mechanisms are the suitable choice. Therefore, if the ratio between AI and human actors shifts within organizations, so too will the need for instrumental accountability mechanisms versus relational mechanisms within organizations. As a result, formal accounting practices should

become central again, in other words, taking a more prominent role in organizations over informal social practices.

Based on this potential shift, this thesis proposes that future research could explore not only which threshold of instrumental accountability should be optimal to adopt for AI systems, but also how different accountability mechanisms could coexist. To be specific, exploring how AI systems operate under an instrumental accountability structure and how human actors operate under and are governed by relational accountability structures. As a result, it should come with the assurance that both technical reliability and moral depth are preserved within organizations. Organizations could namely favor standardization and efficiency within the whole organization over rational depth, thus expecting human actors to meet the same expectations for optimal efficiency again.

Also, it must be concluded that, while this thesis has explored whether AI systems can be held accountable to provide clarity within the complex notion of accountability, it has done so by maintaining the classical structure of accountability as a relationship between the actor, and the counterparty via the account. A relationship encompassing the elements of answerability, responsibility, and sanctionability. This thesis redefines who the actor can be, not the structure itself. This is a consequence of the conceptual choices being made within the methodological framework: to treat AI as an independent actor to the human actor, rather than seeing them interconnected as an augmented hybrid form. This delineation allowed to extend in a clear sense the boundaries of who may be held accountable. In conclusion, it is the practical application that is redefined and broadened, while the abstract notion of accountability remains the same.

Since the abstract notion of accountability remains unaltered, this thesis proposes that future research could instead explore this in depth. In that case, instead of focusing on independent AI processes, the emphasis shifts to the conceptualization of augmented human actors, a hybrid interconnected form of the human and AI. Consequently, this could redefine the abstract notion of accountability. Namely, the current model of accountability assumes that the originator of the action must be individually attributable, which is classically necessary for holding one to account. Therefore, when AI and humans are treated as interconnected, this clarity begins to erode. As a result, the classical concept of accountability, one created in an era pre-AI, may need to be redefined for current times.

At last, returning to the citation generated by Microsoft's AI system, Copilot, it is particularly striking that this thesis, written from a human actor perspective, apparently concludes that it has more trust in (future) AI capacities than AI itself. Perhaps, after all, we can at least conclude that AI does not yet know confidence as an emotional state, which an accountant still probably needs.

7. Appendices

Statement AI

No Generative AI tools were used during the development and writing of this thesis. By submitting this thesis, I declare that I am fully responsible for the accuracy and completeness of its content.

However, Microsoft Copilot has been used once time (on June 28, 2025), to generate a citation in the introduction to this thesis. This was done purely to present an interesting fact, without using it as further evidence for any argument.

Query: "Please answer very concise: Can an AI system be held (morally) accountable with an organization?"

Output: "No, AI systems can't be morally accountable—they lack consciousness and intent. Responsibility lies with the organizations that design, deploy, and oversee them."

How it was used: The output is used as a direct citation in the introduction to illustrate, as an interesting observation, how AI thinks about its own capacities with regard to (moral) accountability.

References

- Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 101805.
- Allen, C., Varner, G., & Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3), 251–261.
- Anscombe, G.E.M. (1957a). Intention. *Proceedings of the Aristotelian Society*, 57, 321–332.
- Anscombe, G.E.M. (1957b). *Intention*. Harvard University Press.
- Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society*, 35(3), 611–623.
- Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2020). An Introduction to Ethics in Robotics and AI. In *Springer Briefs in Ethics*.
- Behdadi, D., & Munthe, C. (2020). A normative approach to artificial moral agency. *Minds and Machines*, 30(2), 195–218.
- Boritz, J.M., Stratopoulos, C.T. (2023). AI and the Accounting Profession: Views from Industry and Academia. *Journal of Information Systems*. 37 (3): 1–9.
- Bovens, M. (1998). *The quest for responsibility: accountability and citizenship in complex organizations*. Cambridge, UK: Cambridge University Press.
- Bovens, M. (2007). *Analysing and Assessing Accountability: a Conceptual framework*. *European Law Journal*, 13(4), 447–468.
- Bovens, M. (2010). Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism. *West European Politics*, 33(5), 946–967.
- Butler, J. (2005). *Giving an Account of Oneself*.
- Cao, Y., & Zhang, W. (2025). How AI is shaping accounting and finance. *The British Accounting Review*, 101650.
- Cheng, L., & Chouldechova, A. (2023). Overcoming Algorithm Aversion: A Comparison between Process and Outcome Control. *CHI'23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–27.
- Chukwuani, V. N., & Egiyi, M. A. (2020). Automation of Accounting Processes: Impact of Artificial Intelligence. *International Journal of Research and Innovation in Social Science (IJRISS)*, 4, 444-449

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, 12(3), 235–241.

Dennett, D.C. (1969). *Content and Consciousness*. Routledge & Kegan Paul.

Dennett, D.C. (1984). *The elbow room: the varieties of free will worth wanting intentional stance*. MIT Press.

Dennett, D.C. (1987). *The intentional stance*. MIT Press.

Dennett, D.C. (1988). The intentional stance in theory and practice. In R. W. Byrne & A. Whiten (Eds.), *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans* (pp. 180–202). Clarendon Press/Oxford University Press.

Dennett, D.C. (1991). *Consciousness Explained*, Penguin Press.

Dennett, D.C. (2009). Intentional systems theory. *The Oxford handbook of philosophy of mind*, Beckermann, A., McLaughlin, B.P., & Walter, S. (Eds). 339–350.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal Of Experimental Psychology General*, 144(1), 114–126. ed. Upper Saddle River, N.J.: Prentice Hall.

Dignum, V. (2022). Responsibility and Artificial Intelligence, in Justin B. Bullock, and others (eds), *The Oxford Handbook of AI Governance*, Oxford Handbooks.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S. J., O’Brien, D., Scott, K., Shieber, S., Waldo, J., Weinberger, D., Weller, A., & Wood, A. (2017). Accountability of AI under the Law: The role of explanation. *SSRN Electronic Journal*.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2022). Explainable AI (XAI): core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9), 1–33.

European Commission. (2021, June 13). Artificial Intelligence Act, Article 14: Human oversight. <https://artificialintelligenceact.eu/article/14/>

Floridi, L., & Sanders, J. (2004). On the Morality of Artificial Agents. *Minds and Machines*, 14(3), 349–379.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707.

Frankfurt, H. (1976). Identification and externality. In University of California Press eBooks (pp. 239–252).

Frankfurt, H. G. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, 68(1), 5.

Friedman, M. (1953). *The Methodology of Positive Economics*. In *Essays in Positive Economics*. University of Chicago Press. pp. 3-43.

Gips, J. (2011). Towards the ethical robot. In *Cambridge University Press eBooks* (pp. 244–253).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hartmann, F. G. H. (2023). *Accounting for action: The possibility of knowledge-based agency*. Radboud University.

Hou, Y. T., & Jung, M. F. (2021). Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2), 1–25.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, 8(4), 195–204.

Jones, G.W. (1992). 'The search for local accountability', pp. 49–78 in S. Leach (ed.), *Strengthening local government in the 1990s*. London: Longman.

Joseph, S. A., Kolade, T. M., Obioha-Val, O., Adebisi, O. O., Ogungbemi, O. S., & Olaniyi, O. O. (2024). AI-Powered Information Governance: Balancing Automation and Human Oversight for Optimal Organization Productivity. *Asian Journal Of Research in Computer Science*, 17(10), 110–131.

Kaplan, A., & Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land?

Keasey, K. & Wright, M. (1993), *Issues in Corporate Accountability and Governance: An Editorial*, *Accounting and Business Research* 23(91A), 291–303.

Kokina, J., & Davenport, T. H. (2017). The Emergence of Artificial Intelligence: How Automation is Changing Auditing. *Journal of Emerging Technologies in Accounting*, 14(1), 115–122.

Kroll, J. A. (2020). *Accountability in Computer Systems*. *The Oxford Handbook of the Ethics of AI*. Dubber, Markus D., Frank Pasquale, and Sunit Das, Eds. Oxford University Press. 2020.

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., Robinson, D. G., and Yu, H. (2016). *Accountable Algorithms*. *University of Pennsylvania Law Review*, Vol. 165, 2017 Forthcoming, *Fordham Law Legal Studies Research Paper No. 2765268*.

Lechterman, T. M. (2022). 'The Concept of Accountability in AI Ethics and Governance', in Justin B. Bullock, and others (eds), *The Oxford Handbook of AI Governance*, Oxford Handbooks.

Leitner-Hanetseder, S., Lehner, O. M., Eisl, C., & Forstenlechner, C. (2021). A profession in transition: actors, tasks and roles in AI-based accounting. *Journal of Applied Accounting Research*, 22(3), 539–556.

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. *Psychological Bulletin*, 125(2), 255–275.

Loi, M., & Spielkamp, M. (2021). Towards Accountability in the Use of Artificial Intelligence for Public Administrations. *AIES'21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 757 – 766.

May, E. (1946). The logos in the Old Testament. *The Catholic Biblical Quarterly*, 8(4), 438–447.

Meng, X. (2023). Data science and engineering with human in the loop, behind the loop, and above the loop. *Harvard Data Science Review*, 5(2).

Messner, M. (2009). The limits of accountability. *Accounting Organizations And Society*, 34(8), 918–938.

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.

Moor, J. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21.

Moss, J. (2014). Right reason in Plato and Aristotle: On the meaning of logos. *Phronesis*, 59(3), 181–230.

Mulgan, R. (2000). 'Accountability': an Ever-Expanding concept? *Public Administration*, 78(3), 555–573.

Odonkor, N. B., Kaggwa, N. S., Uwaoma, N. P. U., Hassan, N. A. O., & Farayola, N. O. A. (2024). The impact of AI on accounting practices: A review: Exploring how artificial intelligence is transforming traditional accounting methods and financial reporting. *World Journal of Advanced Research and Reviews*, 21(1), 172–188.

Poole, D. L., & Mackworth, A. K. (2010). *Artificial intelligence: Foundations of computational agents*. Cambridge University Press.

Raja, A. K., & Zhou, J. (2023). AI Accountability: Approaches, Affecting Factors, and Challenges. *Computer*, 56(4), 61–70.

- Reddy, G. P., & Kumar, Y. V. P. (2023). Explainable AI (XAI): explained. *School of Electronics Engineering*, 1–6.
- Roberts, J. (1991). The possibilities of accountability. *Accounting Organizations and Society*, 16(4), 355–368.
- Roberts, J. (2009). No one is perfect: The limits of transparency and an ethic for ‘intelligent’ accountability. *Accounting Organizations and Society*, 34(8), 957–970.
- Russell, S. & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. 3rd Edition, Prentice-Hall, Upper Saddle River.
- Schweitzer, B. (2024). Artificial intelligence (AI) Ethics in Accounting. *Journal of Accounting Ethics & Public Policy*, 25(1).
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral And Brain Sciences*, 3(3), 417–424.
- Shearer, T. (2002). Ethics and accountability: from the for-itself to the for-the-other. *Accounting Organizations and Society*, 27(6), 541–573.
- Simon, H.A. (1990). Bounded Rationality. In: Eatwell, J., Milgate, M., Newman, P. (eds) *Utility and Probability*. The New Palgrave. Palgrave Macmillan, London.
- Sinclair, A. (1995). The chameleon of accountability: Forms and discourses. *Accounting Organizations and Society*, 20(2–3), 219–237.
- Sparrow, R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1), 62–77.
- Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology*, 8(4), 205–213.
- Stenning, P. C. (1995). *Accountability for Criminal Justice*. In University of Toronto Press eBooks.
- Sullins, J. P. S. (2006). When is a robot a moral agent? *The International Review of Information Ethics*, 6, 23–30.
- Sunstein, C., & Gaffe, J. (2024). An Anatomy of Algorithm Aversion. *Science And Technology Law Review*, 26(1).
- Vosselman, E. G. J. (2016). Accounting, accountability and ethics in public sector organizations: towards a duality between instrumental accountability and relational response-ability. *Administration and Society*, 48(5), 602-627.