A Reexamination of Design- and Policy-oriented Computer Ethics:

Accounting for Increasingly Complex Actor Constellations and Regulatory Developments

Cumulative dissertation with the aim of achieving a doctoral degree

at the Faculty of Mathematics, Informatics, and Natural Sciences

Department of Informatics

Universität Hamburg

submitted by

Mattis Jacobs

2023

Hamburg

Date of Disputation 11. Juli 2023

First Evaluator Prof. Dr. Judith Simon

Second Evaluator Prof. Dr. Ingrid Schirmer

Table of Contents

Abstract

As computer systems become more pervasive in society, computer ethics gains importance. Yet, various approaches to computer ethics face challenges due to 1) increasingly complex actor constellations in many socio-technical systems, which involve a high degree of distributed agency and power, and 2) recent regulatory developments. This thesis reexamines computer ethics in light of these challenges. The focus is on two approaches to computer ethics: first, *design-oriented* computer ethics, dealing with the design of computer technology itself (i.e., separate from the behavior of the developers and users); secondly, *policy-oriented* computer ethics, which aims at formulating and justifying policies (i.e., practices, principles, laws, and rules) for the ethical use of computer technology.

This thesis addresses two main research questions: First, what challenges to the application of different approaches to computer ethics might arise from the social, legal, and technical environment in which computer technology is embedded? Secondly, how can the social, legal, and technical environment be shaped in such a way that approaches to computer ethics can be applied effectively? The thesis develops a perspective on computer ethics that considers shaping the conditions under which approaches to computer ethics can be successfully applied as an additional task of the discipline.

Furthermore, the thesis provides tangible guidance for addressing challenges to policy- and design-oriented approaches to computer ethics and outlines novel opportunities for them. It shows that computer ethics can be used to shape how power manifests among the actors involved in socio-technical systems. In doing so, computer ethics can enable specific actors to apply its approaches more effectively in the future. Moreover, the thesis emphasizes the importance of considering differences in the ability to influence design decisions (even against resistance) among the actors involved in a socio-technical system. Such considerations make it possible to identify powerful individual actors. These can be encouraged (or forced) to shape the socio-technical system in accordance with specific ethical values or principles.

The thesis was written cumulatively and consists of four peer-reviewed research articles. It is based primarily on a critical review and discussion of scholarly literature in philosophy, computer science, and related disciplines. It explores the research questions using examples of blockchain-based systems, platform ecosystems, and artificial intelligence systems. The thesis builds on approaches such as *Value Sensitive Design* and *Disclosive Computer Ethics*. Furthermore, it is guided by a perspective on computer ethics developed by James Moor in his seminal article *What is computer ethics?*, which places particular emphasis on the role of computer ethics in filling conceptual and policy vacuums.

Keywords: Computer Ethics, Value Sensitive Design, Blockchain, Platform Ecosystems, Artificial Intelligence

Kurzfassung

Mit der zunehmenden Verbreitung von Computersystemen in der Gesellschaft gewinnt die Computerethik an Bedeutung. Verschiedene Ansätze der Computerethik stehen jedoch vor Herausforderungen. Diese ergeben sich aus 1) zunehmend komplexer werdenden Akteurskonstellationen in vielen soziotechnischen Systemen, die ein hohes Maß an verteilter Handlungsfähigkeit und Macht mit sich bringen und 2) aktuellen regulatorischen Entwicklungen. In dieser Dissertation wird die Computerethik im Lichte dieser Herausforderungen untersucht. Der Schwerpunkt liegt dabei auf zwei Ansätzen der Computerethik: Erstens der *designorientierten* Computerethik, die sich mit der Gestaltung von Computertechnologie selbst befasst (d.h. unabhängig vom Verhalten der Entwickler und Nutzer); zweitens der *policy-orientierten*[1] Computerethik, die darauf abzielt, *policies* (d.h. Praktiken, Prinzipien, Gesetze und Regeln) für die ethische Nutzung der Computertechnologie zu formulieren und zu begründen.

Diese Arbeit befasst sich primär mit zwei Forschungsfragen: Erstens: Welche Herausforderungen für die Anwendung verschiedener Ansätze der Computerethik können sich aus dem sozialen, rechtlichen und technischen Umfeld ergeben, in welches eine Computertechnologie eingebettet ist? Und zweitens: Wie kann das soziale, rechtliche und technische Umfeld so gestaltet werden, dass Ansätze der Computerethik effektiv angewendet werden können? Die Dissertation entwickelt eine Perspektive auf die Computerethik, die das Schaffen von Bedingungen, unter denen Ansätze der Computerethik erfolgreich angewendet werden können, als eine zusätzliche Aufgabe der Disziplin betrachtet.

Darüber hinaus gibt sie konkrete Handlungsempfehlungen für die Bewältigung von Herausforderungen, die sich für policy- und designorientierte Ansätze der Computerethik ergeben, und zeigt neue Möglichkeiten für diese Ansätze auf. Sie verdeutlicht, dass die Computerethik dazu genutzt werden kann, die Machtverhältnisse zwischen den an soziotechnischen Systemen beteiligten Akteuren

[1] Der Begriff “policy” wird in der Computerethik weit gefasst. Nach Bynum (2008, p. 29) umfasst er “bestehende Praktiken, Prinzipien, Gesetze und Regeln, die das menschliche Verhalten in dieser Gesellschaft bestimmen” [*“existing practices, principles, laws, and rules that govern human behavior within that society*”]. Da es für diesen weit gefassten “policy”-Begriff keine Entsprechung im Deutschen gibt, wird in dieser Kurzfassung der englische Begriff verwendet.

zu beeinflussen. Auf diese Weise kann die Computerethik bestimmte Akteure in die Lage versetzen, Ansätze der Computerethik in Zukunft effektiver anzuwenden. Des Weiteren unterstreicht die Dissertation die Bedeutung einer differenzierten Betrachtung davon, welche Möglichkeiten zur Beeinflussung von Designentscheidungen die an einem soziotechnischen System beteiligten Akteure haben. Solche Überlegungen ermöglichen es, mächtige Einzelakteure zu identifizieren. Diese können dazu ermutigt (oder gezwungen) werden, das jeweilige soziotechnische System im Einklang mit bestimmten ethischen Werten oder Prinzipien zu gestalten.

Die Dissertation wurde kumulativ verfasst und besteht aus vier von Peer-Reviewern begutachteten und in Fachzeitschriften veröffentlichten Artikeln. Sie basiert in erster Linie auf der kritischen Analyse und Diskussion von wissenschaftlicher Literatur aus den Bereichen Philosophie, Informatik und verwandten Disziplinen. Sie untersucht die Forschungsfragen anhand von Beispielen von Blockchain-basierten Systemen, Plattform-Ökosystemen und Systemen der künstlichen Intelligenz. Die Arbeit stützt sich auf Ansätze wie *Value Sensitive Design* und *Disclosive Computer Ethics*. Darüber hinaus orientiert sie sich an einer Perspektive auf die Computerethik, die James Moor in seinem Artikel *What is computer ethics?* entwickelt hat und die insbesondere die Rolle der Computerethik bei der Adressierung von *conceptual vacuums* und *policy vacuums* betont.

Keywords: Computerethik, Value Sensitive Design, Blockchain, Plattform Ökosysteme, Künstliche Intelligenz

Acknowledgments

I would like to express my deep gratitude and appreciation to all those who have supported me throughout my PhD journey.

First and foremost, I am grateful to my supervisor, Judith Simon, for her support and guidance. In particular, I would like to thank her for providing valuable insights and critical feedback while encouraging me to follow my interests and supporting me in realizing my ideas. I would also like to express my special gratitude for her continued close support and mentorship and our productive collaboration on research articles after I left the *Ethics in Information Technology (EIT)* research group as a research associate.

Next, I would like to thank Ingrid Schirmer for her kind offer to serve as my second examiner.

Further thanks go to the project teams of the *Information Governance Technologies (IGT)* and *Governance of and by Algorithms – About algorithmic behaviour control and artificial intelligence (GOAL)* research projects for the rewarding collaboration. Working on both projects helped me to develop my thesis's research questions. Of these project partners, I would especially like to thank my co-authors Christian Kurtz, Judith Simon, and Tilo Böhmann for the productive collaboration on the article *Value Sensitive Design and power in socio-technical ecosystems*, the results of which remained of great importance for my further research.

Moreover, I would like to thank my former colleagues at the *EIT* research group, Laura Fichtner, Pak-Hang Wong, Gernot Rieder, Catharina Rudschies, Ingrid Schneider, Anja Peckmann, and Jason Branford, who provided me with a stimulating intellectual environment. Their feedback, critique, and support have been invaluable to my research. I have enjoyed working with all involved. I am particularly grateful for maintaining such a positive and enjoyable working atmosphere despite the challenges posed by the pandemic.

I am also indebted to my partner, family, and friends, who provided critical feedback, personal support, patience, and encouragement throughout this journey. In particular, I would like to thank Hannah Strothmann, Hannah Jacobs, Lucas Jacobs, Stephan

Jacobs, Mechthild Jacobs, Shushanik Margaryan, Jens Schnitzler, Jonathan Lemke, Jonas Bunsen, and Dafna Burema.

Finally, I would like to acknowledge and express my gratitude for the financial support provided by the Hamburg Ministry of Science, Research and Equality (project *IGT* under the reference LFF-FV 34), the German Federal Ministry of Education and Research (project *GOAL* under the reference 01IS19020), and the DFG (German Research Foundation) (project Science of Intelligence /– EXC 2002/1 "Science of Intelligence" – under the reference 390523135). The financial support has been instrumental in enabling me to conduct this research. Furthermore, I would like to thank the project DEAL for enabling Open Access publishing.

List of Abbreviations

ADM	Algorithmic Decision Making
AI	Artificial Intelligence
EC	European Commission
EU	European Union
IGT	Information Governance Technologies (project)
GOAL	Governance of and by Algorithms (project)
MAC	Media Access Control
NLP	Natural Language Processing
MiCA	Markets in Crypto-Assets Regulation
ML	Machine Learning
VSD	Value Sensitive Design

Eidesstattliche Versicherung/Declaration on Oath

Hiermit erkläre ich,

Mattis Jacobs, geboren am 09. Juli 1989 in Aachen,

an Eides statt, dass ich die vorliegende Dissertationsschrift

"A Reexamination of Design- and Policy-oriented Computer Ethics: Accounting for Increasingly Complex Actor Constellations and Regulatory Developments"

selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

____

Mattis Jacobs

1. Introduction

1.1 Motivation

Computer ethics is a form of applied ethics that addresses moral questions relating to computer technology. It deals with computer systems and practices relating to these systems, such as their development and use (van den Hoven, 2008, p. 49). Since computer technology constantly evolves and related practices change, the subject of computer ethics is a dynamic one. Computer ethics needs to take this evolution of computer technology and related practices into account and refine its concepts and methods accordingly.

This thesis was written in the context of two research projects in which computer ethics provided one of the core research perspectives: *Information Governance Technologies* (IGT)[1] and *Governance of and by Algorithms – About algorithmic behaviour control and artificial intelligence* (GOAL).[2] In both projects, challenges to applying approaches to computer ethics emerged, which generated the need to reexamine these approaches. These challenges were primarily caused by 1) increasingly complex actor constellations in many socio-technical systems, which involve a high degree of distributed agency and power[3], and 2) recent regulatory developments. In this thesis, socio-technical systems are understood to be systems in which "technological components and social arrangements" are intertwined to the degree that requires considering both parts jointly (Bauer & Herder, 2008, p. 601; see

[1] IGT was a project of a research collaboration between the University of Hamburg, the Technical University of Hamburg-Harburg, and the Leibniz Institute for Media Research | Hans Bredow Institute. It focused on human-centered technical possibilities for shaping the digital society with a particular focus on decision-making rights and responsibilities regarding the collection and processing of information.

[2] GOAL was a project of a research collaboration between the Westfälische Wilhelms-Universität, the Karlsruher Institut für Technologie, the Ruhr-Universität Bochum, the University of Hamburg, and the Technical University of Kaiserslautern. It focused on how the governance of risk-prone algorithms can be designed, on the one hand, and explored how algorithms themselves can perform governance functions, on the other hand.

[3] The thesis' definition of the term 'power' is discussed in-depth in Chapter 2.1.4 as well as in Publictions 2 and 4.

also Ropohl, 1999).[4] Examples of such systems discussed in this thesis are AI systems, platform ecosystems, and blockchain-based systems.

The relationship between regulation and computer ethics, and the role of power in computer ethics, are not new issues per se. Yet, some fundamental shifts have occurred in both areas in recent years.

First, there is an increasing shift from 'soft policies' such as ethics guidelines, which "are not legally binding but persuasive in nature" to "so-called hard law—that is, legally binding regulations passed by the legislatures to define permitted or prohibited conduct" (Jobin et al., 2019, p. 389). In the case of the European Union (EU), for example, the General Data Protection Regulation was implemented in 2018 (European Commission, 2016). In addition, far-reaching regulatory proposals such as the Digital Markets Act, the Digital Services Act, and the Artificial Intelligence (AI) Act are either expected to be implemented soon or have been implemented and will be applicable soon (European Commission, 2020b, 2020c, 2021c). Among other things, many of these regulations aim at upholding ethical values and principles. Yet, as a side-effect, these legally binding regulations potentially limit the ability of developers, operators, and users of computer systems to negotiate and act on what they consider to be ethical behavior and design.

Secondly, computer ethics is increasingly shifting its attention to more complex socio-technical systems, such as platform ecosystems, AI systems, and blockchain-based systems. In such systems, the coordination among actors involved in design processes is more difficult than in the design processes of stand-alone, monolithic applications. This is partly due to the fact that power-related issues become more salient with more complex constellations of actors. Consequently, the question of how computer ethics can account for power becomes increasingly pressing. As Friedman et al. (2021) note, this question so far has not been adequately addressed in many areas of computer ethics. Yet, as the research in IGT and GOAL shows, the applicability of some approaches to computer ethics can be called into question due to power-related issues. In order to maintain the applicability of these approaches in the context of complex

[4] This definition will be discussed in more depth and distinguished from other definitions in Chapter 2.2.1.

socio-technical systems, computer ethics must therefore determine how to account for power.

1.2 Goals and Ambition

This thesis reexamines computer ethics in light of the advancing regulation of computer technology and the increasingly complex actor constellations in the socio-technical systems that computer ethics deals with. It explores new challenges for computer ethics that these developments pose, as well as new opportunities they offer. The focus is on two approaches to computer ethics: first, *design-oriented* computer ethics, dealing with the design of computer technology itself (i.e., separate from the behavior of the developers and designers); secondly, *policy-oriented* computer ethics, which aims at the "formulation and justification of policies for the ethical use" of computer technology (Moor, 1985, p. 266).

This thesis addresses two main research questions: First, what challenges to the application of different approaches to computer ethics might arise from the social, legal, and technical environments in which computer technology is embedded? It concerns, in van den Hoven's (2008, p. 59) words, the "institutional and material conditions [which] need to be fulfilled" to allow a successful application of computer ethics. Secondly, how can the social, legal, and technical environment be shaped in such a way that approaches to computer ethics can be applied effectively? The thesis develops a perspective on computer ethics that considers shaping the conditions under which approaches to computer ethics can be successfully applied as an additional task of the discipline.

Furthermore, the thesis provides tangible guidance for addressing challenges to policy- and design-oriented approaches to computer ethics and outlines novel opportunities for them. It shows that computer ethics can be used to shape how power manifests among the actors involved in socio-technical systems. In doing so, computer ethics can enable specific actors to apply its approaches more effectively in the future. Moreover, the thesis emphasizes the importance of considering differences in the ability to influence design decisions (even against resistance) among the actors involved in a socio-technical system. Such considerations allow identifying powerful individual actors who can be encouraged (or forced) to shape the socio-technical system in accordance with specific ethical values or principles.

1.3 Structure of the Thesis

The thesis was written cumulatively and is based on four peer-reviewed research articles. It is structured as follows. Part 1 comprises Chapters 1 to 6 of the thesis. These chapters develop the dissertation framework. Part 2 comprises chapters 7 to 10 of the thesis. These chapters contain the published research articles that constitute the thesis' main research contribution.

Following the Introduction, Chapter 2 introduces the theoretical and empirical foundations of this thesis. It provides a basic understanding of computer ethics and the socio-technical systems examined in this thesis. Chapter 3 introduces the methodology of the thesis. It explains the perspective of computer ethics from which the argumentation of the thesis emerges and describes the specific steps that were taken in the research that underlies each of the publications. Chapter 4 provides an overview of the publications that are included in this thesis. Chapter 5 summarizes and links the contributions of these publications and assembles them into an integrated overall picture. Lastly, Chapter 6 provides a final reflection on the thesis. It focuses on research at the intersection of computer science and philosophy, as well as limitations and implications for further research.

Dissertation Framework

Included Publications

2 Theoretical and Empirical Foundations

This chapter outlines the theoretical and empirical foundations of the thesis. Due to the thesis' cumulative approach, some of these foundations have also been addressed in the included publications. Accordingly, the aim of this chapter is twofold. First, it integrates the theoretical and empirical foundations outlined in the included publications into one coherent framework. Secondly, where the format of a research article did not allow for exhaustive explanation or analysis, additional background information is provided.

This chapter proceeds as follows. Chapter 2.1 provides an overview of the theoretical foundations of computer ethics relevant to this thesis. Chapter 2.2 provides background information on the socio-technical systems discussed in this thesis: blockchain-based systems, platform ecosystems, and AI systems. The respective subchapters outline the thesis' understanding of these systems, address why they are appropriate cases for exploring the research questions, and introduce relevant definitions and concepts.

2.1 Introducing Computer Ethics

2.1.1 Historical Overview

Computer ethics has developed over several decades, and perspectives of computer ethics have evolved significantly over time. While the study of the societal and ethical implications of computer technology can be traced back to Wiener's cybernetics and information ethics (Wiener, 1961, 1989), the term "computer ethics" was coined by Walter Maner and his computer ethics initiative (Bynum, 2008). Earlier publications focus primarily on practices relating to computer technology (especially its use) and, at a more abstract level, the challenges to existing ethical concepts (Bynum, 2008; see also Weizenbaum, 1976).

In the seminal paper "What is computer ethics?" Moor develops a theoretical grounding for the discipline. He argues that "***computer ethics*** [emphasis in original] is the analysis of the nature and social impact of computer technology and the corresponding formulation and justification of policies for the ethical use of such technology" (Moor, 1985, p. 266).[5] Furthermore, as discussed more in-depth in Chapter 2.1.2, Moor "went beyond descriptions and examples of computer ethics problems and offered an explanation of why computing technology raised so many ethical questions compared to other technologies" (Bynum, 2008, p. 34).

In the following years, the scope of computer ethics expanded. Especially the role of professional computer specialists such as developers became a research topic (Johnson & Miller, 2009). This new focus reflects in the development of professional codes of ethics and professional conduct for computer scientists and engineers (see, e.g., ACM, 1992). Gotterbarn (1991) emphasizes the role of designers and developers among these professionals, as well as the role of "ethical decisions made during the development" of computer technology.

In line with the design turn in applied ethics, which shifted the focus to the "design of institutions, infrastructure, and technology" (van den Hoven, 2008, p. 58), computer ethics also began to consider the design of computer technology as a subject of the discipline.[6] Disclosive Computer Ethics (Brey, 2000, 2010; Introna, 2005) and Value Sensitive Design (Friedman et al., 2008; Friedman & Hendry, 2019) are indicative approaches for the turn to this new category of problems.

Disclosive Computer Ethics focuses "on morally opaque practices" (Brey, 2010, p. 51) and "moral deciphering of computer technology" or, more specifically, its "design features" (Brey, 2000, pp. 11–12). In contrast, Value Sensitive Design emphasizes that moral features in the design of (computer) technology can not only be analyzed *ex-post* but can also be taken into account in design processes. It "provides theory, method, and practice to account for human values in a principled and systematic manner throughout the technical design process" (Friedman & Hendry, 2019, pp. 3–4).

2.1.2 Theoretical Foundations of Policy-oriented Computer Ethics

The importance of regulating computer systems stems from the fact that novel computer technologies allow their users to perform new activities (or old activities in new ways).

[5] Accordingly, in this thesis, such approaches are labelled "policy-oriented."

[6] Accordingly, in this thesis, such approaches are labelled "design-oriented."

This circumstance can lead to situations “in which we do not have adequate policies in place to guide” our actions (Moor, 2005, p. 115).

Publication 4 elaborates on this as follows:

Publication 4

> “[…] given these new choices for actions, often ‘no policies for [ethical] conduct’ exist or ‘existing policies seem inadequate’ (Moor, 1985), as the developments in computer technology outpace ‘ethical, […] and legal developments’ (Floridi & Sanders, 2002).”

While this is true also for other technologies, computer ethicists have pointed out that the “logical malleability” (Moor, 1985) of computer technology makes it a universal tool. Accordingly, it enables human beings to do an “enormous number of new things” (Bynum, 2008, p. 34) for which policies are lacking. Furthermore, in contrast to other technologies which lack the malleability of computer systems, these policy vacuums concern many different application contexts.

The term ‘policy’ is defined very broadly in computer ethics and includes “existing practices, principles, laws, and rules that govern human behavior within that society” (Bynum, 2008, p. 29) or simply “laws, rules, and customs” (Moor, 2005, p. 115). According to Moor, the goal of computer ethics is to develop coherent conceptual frameworks for understanding ethical problems involving computer technology and ultimately to replace such “policy vacuums with good policies supported by reasonable justifications” (Moor, 2001, p. 89).

A central problem for computer ethics is to constructively address “conceptual vacuums” or “conceptual muddles” that often occur in the context of policy vacuums. This, according to Moor, is because computer technology transforms “many human activities and social institutions” and, therefore, addressing problems of computer ethics requires the reevaluation and advancement of ethical and philosophical concepts and theories (Moor, 1985, pp. 270–271).

Moor (1985, pp. 266–267) exemplifies this problem by discussing policies for protecting computer programs as intellectual property. He argues that to develop adequate policies, one has to first answer the question of what a computer program is: an idea (which cannot be owned by anyone), an “*expression* of an idea” which might be protectable by copyright, or “a *process*” that is owned and might be protectable as a patent. More recently, the emergence of the blockchain technology raised similar questions on "the legal nature of cryptocurrency" (Bolotaeva et al., 2019). In particular, there is debate about whether cryptocurrency tokens should be treated as a currency or a commodity, with far-reaching implications for their regulation.

According to Moor (1985, p. 266), developing appropriate policies requires addressing such conceptual vacuums first. Consequently, he argues that "[i]ndeed, much of the important work in computer ethics is devoted to proposing conceptual frameworks for understanding ethical problems involving computer technology."

2.1.3 Theoretical Foundations of Design-oriented Computer Ethics

In contrast to the theoretical foundations of policy-oriented computer ethics, the theoretical foundations of design-oriented computer ethics are based on a more controversial philosophical premise. It holds that not only the human use of computer systems can be ethically evaluated, but also the computer systems themselves. This premise is based on the assumption that technical artifacts are not morally neutral but have tendencies to promote or demote particular moral norms and values (Brey, 2010; Introna, 2005).

Brey (2010) argues that "[t]he notion that technology can have moral properties is an extension of the notion that it can have political properties," as suggested, for instance, by Langdon Winner. Winner (1980, pp. 127–128) states that technological artifacts 'have politics' because they "are ways of building order in our world." In that sense, he argues, technology is similar to "legislative acts or political foundings." Likewise, Latour (1992) argued that tasks can be delegated to technological artifacts and that those artifacts can be used to constrain the actions of (other) actors. The idea "that technological artifacts (and in particular computer systems and software) have built-in tendencies to promote or demote the realization of particular values" (Brey, 2010, p. 43) reflects such perspectives on technology.

Some approaches to design-oriented computer ethics go further. They argue that computer ethics can not only investigate and disclose built-in tendencies of computer systems to promote or demote ethical values. Instead, they claim that computer ethics can also support designing computer systems in ways that ensure that they promote or demote ethical values as intended by their designers. This pragmatic turn in computer ethics advocates the conscious consideration of ethical values as a concrete goal and quality criterion of technical design (Flanagan et al., 2001; see also Friedman & Nissenbaum, 1996; Nissenbaum, 2005). Approaches include, for instance, Value Sensitive Design and Values in Design (Friedman et al., 2002; Friedman et al., 2008; Nissenbaum, 2005; Simon, 2016). They resonate with similar approaches within computer science, such as privacy by design (Cavoukian, 2011; van Rest et al., 2014) and Fair, Transparent, and Accountable ML (FAT-ML, 2018; Lepri et al., 2018).

2.1.4 Philosophy of Power and Information Technology

Computer ethics discusses power in a variety of contexts. These include the way in which (computer) technologies distribute power in society, the power of individual actors in specific socio-technical systems, and the role of power in designing computer systems (see, e.g., Bratteteig & Wagner, 2012; Brey, 2008; Friedman et al., 2021; Nieborg et al., 2020).

To adequately describe and analyze phenomena in such different contexts, computer ethics introduces different notions of power. These notions can be subsumed in two groups: power in terms of "power over" and power in terms of "power to."

Notions of power in terms of "power-over" focus on how one actor can exercise power over another. In the philosophy of technology, notions of power in terms of "power over" are, for instance, applied when discussing how technical artifacts are used "to establish or maintain asymmetrical power relations" of one actor over another (Brey, 2008, p. 88). Here, Foucault's account of power in his discussion of the Panopticon in *Discipline & Punish* (Foucault, 2012) is a prime example. Foucault argues that the architectural features of a particular prison design that allow guards to observe prisoners without them being able to recognize if they are being watched render it a "mechanism of power" (Fontana-Giusti, 2013, p. 89). This is because it gives the observing guards power over the observed prisoners, who exercise self-discipline and self-control in the constant expectation of being watched.[7] Such a 'Foucauldian' notion of power is frequently adopted in computer ethics when discussing power in the context of (digital) surveillance (see, e.g., Jonsson, 2006; Saulles & Horner, 2011).

[7] Please note that this account of Foucault's discussion of the Panopticon is highly abbreviated and far from exhaustive. Here, it merely serves as a prominent example of a notion of power in terms of "power-over."

In contrast, notions of power in terms of “power to” focus on the ability of agents to “realize a certain outcome” (Brey, 2008, p. 75). Such notions of power largely draw on Max Weber, who defined power as the capability in “a social relationship of enforcing one’s own will even against resistance” (Weber, 2019, p. 134). In the philosophy of technology, notions of power in terms of “power to” are, for instance, applied when discussing which actors can influence design decisions of technical artifacts. Since the thesis contributes to this discourse, it adopts this outcome-oriented notion of power accordingly.

To investigate how the power to influence design decisions manifests among the actors involved in socio-technical systems, the thesis adopts a “systemic view” of power that “regards power as the property of broader social, economic, cultural, and political networks, institutions, and structures” (Sattarov, 2019, p. 20). Focusing on socio-technical systems, the thesis also considers their technical features. It analyses how these features “confer differentials of dispositional power on agents, thus structuring possibilities for action” (Haugaard, 2010, p. 425).

As discussed in Publication 4, the discourse in computer ethics relating to power in design processes often adopts a societal perspective. It raises the question of which societal actors are involved in design processes. A central theme is the call for a ‘democratization of technology,’ meaning that design processes “should be arranged to guarantee broad public participation, in which all stakeholders have their voice heard on these processes” (Brey, 2008, p. 92; see also Sclove, 1992; Slota, 2020; Zimmerman, 1995).

In contrast, computer ethics only recently started to emphasize power-related issues in the development processes themselves. More specifically, it concerns only to a lesser degree the questions of how power is distributed among the actors involved in the design process, or, in other words, who has the power to dictate design decisions once the respective voices have been heard.[8] Yet, more recently, accounting for power in design processes has been recognized as a core challenge of design-oriented approaches to computer ethics (Friedman et al., 2021).

[8] However, there are notable exceptions. See, e.g., Bratteteig and Wagner (2012); Shilton (2012); Shilton and Greene (2019). These are subject to discussion later in this thesis.

Furthermore, the thesis refines its perspective on power by incorporating concepts relating to power in specific types of socio-technical systems, such as boundary resources (Eaton et al., 2015; Ghazawneh & Henfridsson, 2013; Karhu et al., 2018) and value levers (Shilton, 2012; Shilton & Greene, 2019). Here it builds not only on philosophical literature, but also on information systems literature. These concepts are introduced in detail in Publication 2. Moreover, they are also discussed in Chapter 5.

2.2 Introducing Socio-technical Systems

As mentioned in the Introduction, this thesis is motivated by challenges to the application of computer ethics that emerged in the IGT and GOAL research projects. Accordingly, the thesis discusses its research questions using the example of socio-technical systems central to these research projects around which these challenges arose. These are blockchain-based systems, platform ecosystems, and AI systems.

The focus on blockchain-based systems and platform ecosystems has roots in the IGT project. One of the core research goals here was to describe the relationship between ethical principles and architectures[9] of IT systems.

In an early project phase, platform ecosystems and blockchain-based systems were selected as cases to investigate some of the project's research questions. The selection was based on the identification of links between features of the respective systems and ethical values and principles. These were mainly *privacy* in the case of platform ecosystems and *trust* in the case of blockchain-based systems. The links were established based on literature reviews, on the one hand, and assessing critical incidents based on media reports, on the other hand.[10]

In contrast, the focus on AI systems has its roots in the GOAL project. This focus was determined in the initial project proposal already. While the project title refers to 'algorithms,' the project outline specifies that it mainly focuses on 'novel types of algorithms' which derive rules from data instead of following 'deterministically

[9] According to IEEE standard 42010:2011, architechtures are "fundamental concepts or properties of a system in its environment embodied in its elements, relationships, and in the principles of its design and evolution."

[10] The process for selecting these cases is described in more detail in Chapter 3.

stringent programmed arithmetic operations.' Specifically, the project focused on systems applying machine learning. In line with recent policy papers and policy proposals by the European Commission, the thesis considers such systems as AI systems (European Commission, 2019, 2020a, 2021c).

Chapter 2.2.1 begins by outlining the thesis' definition and perspective on socio-technical systems. Next, Chapters 2.2.2 to 2.2.4 introduce each system in more detail. Each subchapter provides background information on the respective socio-technical systems (including relevant concepts and definitions) and discusses their relevance to the thesis. Finally, Chapter 2.2.5 discusses the similarities and differences between the systems.

2.2.1 Socio-technical Systems and Ecosystems

The term ‚socio-technical system' is a contested one. Different disciplines define it differently. This thesis' definition of the term builds mainly on the Philosophy of Technology, the Philosophy of Science and Engineering, and Science and Technology Studies. From these perspectives, the idea of socio-technical systems is based on the recognition that many systems "require technical artifacts and social arrangements to function." In such systems, the "technological components and social arrangements" are often intertwined to the degree that requires considering both parts jointly (Bauer & Herder, 2008, p. 601; see also Ropohl, 1999). Correspondingly, van House (2004, p. 5) describes socio-technical systems as systems "consisting of both technology and the social, inseparable, mutually constituted" (van House, 2004, p. 5).[11]

Bauer and Herder (2008, p. 601) acknowledge that developing "an analytically precise definition" that distinguishes between socio-technical and mere technical systems is difficult to make. The question of when "economic, legal and political factors are appropriately treated as exogenous factors" (Kroes et al., 2006, p. 804) and when it is necessary to analyze 'the technical' and 'the social' jointly is difficult to answer in general terms.

[11] This perspective needs to be distinguished from, for example, the usual perspective on socio-technical systems in organizational development and information systems. Here, the socio-technical systems are understood to consist of the technical system, the business/organizational environment, work processes, and the people involved in those processes (Hansen et al., 2019, p. 12).

Kroes et al. (2006, p. 804) emphasize that whether a system is considered a socio-technical system or a mere technical system depends on how the boundaries of a system are drawn. Furthermore, they argue that there may be “pragmatic reasons, depending on the specific problem and purpose at hand, to draw the boundaries such that these [social] factors become either exogenous or intrinsic to the system under consideration.” In that sense, framing a system as a socio-technical system is as much an act of taking a certain perspective on the system as it is a statement about system properties.

Accordingly, Kroes et al. (2006, p. 804) note that this “raises the question of which kind of purposes allow these factors to be treated as exogenous and which as intrinsic to the system and on what grounds.” There are various reasons to consider social factors as a part of systems in ethical assessments and reasoning. To take the example of AI systems, Rieder et al. (2021, p. 30) argue that discussing concepts like ‘trustworthy AI’ only makes sense if AI systems are “framed as *socio-technical* systems that include human agents designing, creating, managing and/or operating them.” This is because trust – at least in a motivation-attributing account that goes beyond game-theoretical considerations – is not applicable to mere technical systems (see also Nickel, 2013). Similarly, Sartori and Theodorou (2022, p. 3) argue that “[c]alls for responsible AI” can only be discussed meaningfully if AI systems are considered socio-technical systems. That is, AI systems need to be understood as “the combination of the technical component (i.e. the code and—if used—the data) and social elements (i.e. the stakeholders responsible for the system and the society in which the system is deployed)” (see also Dignum, 2019, 2020).

This thesis focuses mainly on how ‘the social’ and ‘the technical’ are when designing systems. It suggests that systems should be considered to be socio-technical systems if both dimensions “co-evolve, each enabling and constraining […] the other” (Bauer & Herder, 2008, p. 601) and therefore need to be considered jointly in design processes. A particular focus here is on how social and technical features of a system determine whose values are accounted for in design processes and how these features can be shaped by technology and policies.

Moreover, Publication 2 introduces the term ‘socio-technical ecosystem,’ encompassing platform ecosystems and blockchain-based systems. The term ‘ecosystem’ is used to highlight that both socio-technical systems are “systems of systems" that dynamically evolve rather than having an entirely pre-planned design (cf. Kurtz et al., 2018). The article defines socio-technical ecosystems following McConahy et al. (2012, p. 1) as

Publication 2

> "a dynamic community of competing and interdependent people, organizations, and computing systems operating in a complex, capricious environment."

However, the properties that make a system an ecosystem do not always apply to the systems discussed in subsequent publications. Accordingly, these publications and the dissertation framework focus on socio-technical systems, not ecosystems.

2.2.2 Blockchain-based Systems

The origin of the blockchain technology is a whitepaper titled *Bitcoin: A Peer-to-Peer Electronic Cash System* (Nakamoto 2008), published under the pseudonym "Satoshi Nakamoto" on a cypherpunk mailing list. As outlined in Publication 1, the whitepaper described a novel technical approach to realizing cryptocurrencies that are free of centralized authorities and do not root in incumbent institutions. After the concept's first implementation – resulting in the cryptocurrency *Bitcoin* – the technology's application area quickly expanded. Besides allowing the secure exchange of digital tokens without relying on third-party intermediaries (cf. Swan & Filippi 2017), subsequent blockchain-based systems, such as *Ethereum,* enabled self-executing smart contracts, decentralized applications (DApps), decentralized autonomous organizations (DAOs), and several other novel phenomena and organizational structures (Buterin 2014).

The decentralized nature of blockchain technology and the (alleged) potential for disintermediation of social processes motivates the interest of this thesis (and computer ethics more generally) in the technology.

Technical Background:

Publication 2 describes the blockchain technology on a technical level in more detail. As the thesis as a whole, it focuses on open, permissionless systems:

Publication 2

> “A blockchain is a distributed, encrypted, chronological database of transactions recorded by a distributed network of computers (Morabito, 2017; Wright & De Filippi, 2015). It contains ‘every transaction that has been carried out and shared among those participating in the network’ (Morabito, 2017, p. 4). The entries are ‘encrypted and organized’ in ‘smaller datasets referred to as ‘blocks,’’ each of which references ‘to the preceding block in the blockchain’ (Wright & De Filippi, 2015, p. 7). A consensus mechanism warrants the integrity of each transaction over the network. Contrary to other approaches in computer security, in open, permissionless blockchains, the consensus mechanisms are not based on access control, i.e., on ‘carefully vetting participants and excluding bad actors’ (Antonopoulos, 2014). Instead, they rely on economic incentive systems that aim at motivating actors—referred to as miners (Alsindi & Lotti, 2021)—to participate in the validation process and ensuring that it is ‘more profitable and attractive [for them] to contribute to the network than to attack it’ (Brekke & Alsindi, 2021, p. 2). As a result of this approach, ‘the key characteristics of a blockchain [...] are that it is: distributed, decentralized, public or transparent, time-stamped, persistent, and verifiable.’ (DuPont & Maurer, 2015, p. 2). Moreover, the blockchain technology is not restricted to the record-keeping function it utilises in its origin in cryptocurrencies. More recently developed blockchain-based systems such as Ethereum incorporate Turing-complete virtual machines that allow executing not only simple transactions but also more complex operating steps. In turn, this enables running decentralised second-layer applications (DApps) as services on top of the system.”

To understand how data is structured in the blockchain and which role consensus mechanisms play, it is helpful to look at how data is integrated into a concrete system. The following paragraphs describe this process using the example of the Bitcoin blockchain and its so-called *proof-of-work* consensus mechanism.

In the Bitcoin blockchain, a transaction must first be defined, authenticated, and broadcasted to the network by the sender. Secondly, a node on the network must link the transaction to other pending transactions and broadcast this "block" of new transactions to the network. Thirdly, the block needs to be validated by other nodes in the network and approved as a new block in the existing chain of blocks (Morabito, 2017). The following paragraphs describe each required step in detail.

Transactions are composed by the sender of a transaction (Morabito, 2017). Each transaction contains the following information: 1) the sender's address, 2) the receiver's address, 3) the amount of Bitcoin sent (consisting of the amount that the receiver shall receive plus a transaction fee), and 4) proof of ownership (i.e., a "digital signature […] which can be independently validated" by other nodes in the network) (Antonopoulos, 2017). After composing the transaction, the sender broadcasts it to the other nodes in the network.

Addresses and signatures are generated as public/private key pairs (Swan, 2015). The private key is used to generate the public key, which serves as an address for transactions, using a one-way hash function. Subsequently, the private key can be used to generate digital signatures to prove ownership of the address and the funds deposited at that address (Antonopoulos, 2017).

When receiving transactions, nodes can validate the transactions by checking their digital signature and the funds deposited at the sending addresses. Which funds are deposited at which addresses can be inferred from publicly available data in the blockchain. If the digital signature is deemed valid and the required funds are deposited at the sender's address, nodes in the blockchain network can integrate it into a block with other pending transactions.

A block in the Bitcoin blockchain consist of 1) a hash value representing the previous block in the chain, 2) a number of (hashed) pending transactions which are summarized in Merkle tree[12] to produce "an overall digital fingerprint of the entire set of transactions" in the block (Antonopoulos, 2017, p. 202), a random number called "nonce," which is used in the proof-of work consensus mechanism, and a timestamp.

However, a block must meet certain criteria in order to be acceptable according to the criteria of the consensus mechanism. In the case of the Bitcoin blockchain, there must be a preset number of leading zeros in the hex representation of the block hash (i.e.,

[12] "A Merkle tree is constructed by recursively hashing pairs of nodes until there is only one hash, called the root, or merkle root" (Antonopoulos, 2017, p. 202).

the hash value of the concatenation of the strings 1) - 4) above). As “the result of a hash function is virtually unpredictable and irreversible, the only way to validate a block is to try repeatedly, randomly modifying the nonce value until a hash matching the specific target appears by chance” (Dos Santos, 2017, p. 622).

Using brute force to find a nonce that meets the above criteria requires investing computing power. The difficulty of finding a nonce (and thus the on average amount of required computing power) that makes the result of the hash function meet the target requirements can be adapted by changing the required number of leading zeros in the hex representation of the block hash.

Solving this ‘puzzle’ to find nonce meeting the target requirements and thereby creating a new valid block is called mining. While mining is costly in computing power (and thus electricity and hardware), engaging in it can be worthwhile due to the economic incentive system underlying the Bitcoin blockchain. The miner who first successfully finds a nonce that meets the target requirements is rewarded with a block reward that is predefined in the consensus protocol as well as the transaction fees determined by the senders of the transactions. These newly generated funds are credited to the miner’s address.

However, if a miner proposes a block containing either invalid signatures or funds not possessed by the senders, the network ignores the proposed block and its transactions. In that case, the miner would have invested (computing) resources without gaining a reward.[13] With this economic incentive system, Bitcoin attempts to make contributing to the network “more profitable and attractive […] than to attack it (Brekke & Alsindi, 2021).

Involved Actors:

Within the Bitcoin network, different types of nodes fulfill different functions. Mainly, these functions are wallet services, mining, holding a copy of the full blockchain, and network routing (Antonopoulos, 2017). The following paragraphs explain these functions and the different types of nodes that exist in the Bitcoin network.

[13] Note that no similar amount of computing (or other) resources are required to verify the validity of a block.

In the case of Bitcoin, by default, all nodes include the routing function, i.e., they "validate and propagate transactions and [new] blocks, and discover and maintain connections to peers" (Antonopoulos, 2017; see also Werbach, 2018).

In contrast, only some nodes (called "full nodes") administer a "complete and up-to-date copy of the blockchain" (Antonopoulos, 2017, p. 172). This means that they store the entire transaction history of the Bitcoin network, which allows them to verify that transactions are valid. This is because the transaction history allows them to determine whether the sender of a transaction actually owns the funds they intend to send.

Light nodes (also known as Simplified Payment Verification (SPV) nodes), on the other hand, retain only the limited data necessary to operate. They are "designed to run on space- and power-constrained devices, such as smartphones, tablets, or embedded systems" and rely "on peers to provide partial views of relevant parts of the blockchain on demand" (Antonopoulos, 2017, p. 183). Thus, they are useful for users who want to make transactions but do not have the resources to run a full node.

Both full nodes and SPV nodes provide wallet services. A wallet "allows users to manage a collection of private keys corresponding to their accounts and to create and sign transactions on the Bitcoin network" (Xu et al., 2019, p. 34).

Miners are nodes that engage in the mining process described further above. Some miners also maintain a full node, while others use specialized lightweight client software to participate in cooperative (or: pool) mining. Users of such lightweight client software rely "on a pool server to maintain the full node" (Antonopoulos, 2017, p. 173; see also Xu et al., 2019, p. 30)

Furthermore, some nodes provide functions that are not defined within a blockchain's protocol but exist on the respective system's periphery. These are, for instance, cryptocurrency exchanges. Most cryptocurrency exchanges offer two kinds of exchanges: 1) the exchange of regular currencies (i.e., fiat money issued by a government) and cryptocurrency tokens, and 2) the exchange of cryptocurrency tokens of different blockchain-based systems (De Filippi & Wright, 2018). In many cases, users also rely on cryptocurrency exchanges to administer their wallets.

The high level of decentralization in open, permissionless blockchains has some drawbacks. These mainly concern the performance of these systems. Compared to more centralized systems, keeping a copy of all transactions on a multitude of nodes instead of relying on one centrally administered copy creates a huge overhead. That is, "it requires high redundancy in its messaging" (Szabo, 2017). In addition, the consensus mechanisms that ensure the integrity of a blockchain are often based on economic incentive systems that require high resource inputs, such as computing power and, consequently, energy (Alsindi & Lotti, 2021). Thus, the way in which blockchain-based systems maintain integrity in an open and decentralized manner comes at the cost of being "inefficient by design" (Smith, 2017, p. 2301). For a given purpose, these disadvantages must be weighed against the advantages of using a blockchain-based system.[14]

Research Interest:

From a computer ethics perspective, open, permissionless blockchains are a relevant subject as they utilize computer technology to enable novel types of societal interactions. While the hype around the technology has somewhat subsided, some researchers predicted great disruptive potential for the technology a few years ago. The technology was said to have the potential to "to transform political institutions that are central to contemporary human societies, such as money, property rights regimes, and systems of democratic governance" (Reijers et al., 2016, p. 134). These projections were also met with much criticism. Overall, the appeals to such philosophical themes in the evaluation of technology have provoked (and to some extent still provoke) an intense debate in computer ethics.

Within these debates, which role trust plays in blockchain-based systems emerged as a core and highly contested question. As Publication 1 notes

Publication 1

> "[...] the role that trust plays in these systems is understood and portrayed in various manners. The blockchain technology is said to enable (Underwood, 2016, p. 16) and establish (Krishna, 2015) trust as well as to redirect it (Werbach, 2018, p. 30), to substitute for it (Freeman et al., 2020, p. 69), and to make it obsolete (Nakamoto, 2008, p. 8). Furthermore, there is disagreement on whom or what users have to trust when using the

[14] Due to these disadvantages, there are also alternative approaches to blockchain-based systems (private blockchains, consortium blockchains) that are less decentralized and therefore less susceptible to the outlined problems. However, the interesting features of blockchain-based systems from the perspective of computer ethics are based precisely on their decentralized nature and the resulting potential for disintermediation. Therefore, these alternative approaches are not discussed in this thesis.

> blockchain technology: (only) code, math, algorithms, and machines (Maurer et al., 2013; Nakamoto, 2008), or still (also) human actors (Botsman, 2017; Walch, 2019b; Werbach, 2018)."

Publication 1 investigates the role that trust plays in blockchain-based systems. The paper outlines how the lack of a shared understanding of the term "trust" leads to diverging interpretations of the technology's core features. Publication 2, on the other hand, approaches blockchain-based systems from the perspective of VSD. It examines how actors involved in blockchain-based systems can influence design decisions on the system's protocol, which sets the rules by which actors involved in the system interact.

2.2.3 Platform Ecosystems

Nowadays, a vast proportion of digital interactions is facilitated by digital platforms (cf. Lusch & Nambisan, 2015). This gives platforms social relevance and makes them an important research object of computer ethics and related disciplines. According to Nieborg et al. (2020, p. 1), the way that platforms shape information flows disrupts "societal institutions and industry sectors" while potentially conflicting with "vital public values" and undermining "socioeconomic equality, democratic processes, and the quality of public services" (see also van Dijck et al., 2018). Furthermore, within platform ecosystems, there are large power imbalances between different groups of actors, which is also a topic of discussion in computer ethics (Hestres, 2013; Shilton & Greene, 2016, 2019).

Yet, the term "platform" itself is a contested one. It varies in different disciplines and evolved over time (Baldwin & Woodard, 2009; Poell et al., 2019). In its definition of the term 'platform,' this thesis builds mainly on Baldwin and Woodard (2009), Reuver et al. (2018), and Tiwana et al. (2010). It defines platforms as

Publication 2

> "a software-based system as the core with an extensible codebase that enables functionality for users through additional software subsystems in the form of peripheral applications – or modules – that interoperate with it."

Technical Background:

According to Baldwin and Woodard (2009, p. 19), “[p]roduct and system designers have long exploited opportunities to create families of complex artifacts by developing and recombining modular components.” Platforms, they argue, are one frequently used design pattern in which “a set of stable components [...] supports variety and evolvability in a system.”

In this design pattern, peripheral applications or modules make (shared) use of the existing and stable components which provide the “core functionality” of the platform (Tiwana et al., 2010, p. 676). Thereby, such modules can be created and launched efficiently. Conversely, by incorporating such modules or “add-on software subsystems” (Tiwana et al., 2010, p. 675), the platform extends its functionality. Mostly, external actors develop and operate these modules (Reuver et al., 2018).

To enable interactions between the platform’s core functions and modules, platforms provide interfaces through which modules can interact with the platform. By providing these interfaces, the platform provider also sets “[s]pecifications and design rules that describe how the platform and modules interact and exchange information” (Tiwana et al., 2010, p. 676). As discussed later in the thesis, platform providers can use the design of interfaces to influence the design decisions of developers of modules.

Involved Actors:

Publication 2 introduces the central actors involved in platform ecosystems as follows:

Publication 2

> “In platform-based ecosystems, mainly actors of four different groups come together: users, platform providers, app providers, and third parties.”

The constellation of actors in platform ecosystems is best understood in terms of two-sided markets. As Hein et al. (2020, p. 91) note, platforms create these markets “via the orchestration of transactions” between users and service providers. Platform providers provide the core functionality and interfaces to integrate additional modules. Conversely, developers of modules provide additional functionality, services, and content on the platform that users can access via the platform (Constantinides et al., 2018; Ghazawneh & Henfridsson, 2013).

Lastly, in many platform ecosystems, developers of modules integrate software elements or services provided by third parties. Yet, this thesis focuses mainly on platform providers and developers of modules. This is because studies on platform ecosystems have paid less attention to the role of this group of actors (Kurtz et al., 2022). Accordingly, there was less empirical material on the role of these actors to analyze and interpret from a computer ethics perspective.

Research Interest:

Due to their increasingly important role in facilitating digital interactions in various societal contexts, platform ecosystems are becoming a crucial research subject of computer ethics. As other (socio-)technical systems, "platforms are neither neutral nor value-free constructs; they come with specific norms and values inscribed in their architectures [and] these norms may or may not clash with values engraved in the societal structures in which platforms vie to become (or are already) implemented" (van Dijck et al., 2018, p. 3). Computer ethics can assist in analyzing and designing these architectures and impact how platforms facilitate digital interactions.

In addition, literature in computer ethics and related disciplines has critically examined the role of platform providers (Hestres, 2013; Nieborg et al., 2020). Within their respective ecosystems, platform providers control access to the platform and its resources. This raises the question of the extent to which platform providers can interfere with the design choices of developers of modules. This, in turn, has implications for the applicability of design-oriented approaches to computer ethics.

The critical perspective on platforms and platform providers aligns with recent debates in the regulatory domain. Within the European Union, especially the recently introduced Digital Markets Act (European Commission, 2020c) addresses the power of platform providers in platform ecosystems. Publication 2 discusses several requirements of the draft[15] of this regulation that aim at curbing problematic exploitations of power by platform providers. These include

[15] When Publication 2 was published, the Digital Markets Acts was not signed into law yet.

Publication 2

> "[...] the proposed requirements for gatekeepers to 'allow business partners,' such as the providers of peripheral applications in a platform-based ecosystem, 'to offer the same products or services to end users through third party online intermediation services at prices or conditions that are different from those offered through the online intermediation services of the gatekeeper'[16] (European Commission, 2020c, art. 5 (b)) or to 'provide effective portability of data generated through the activity of a business user or end user [...]' (European Commission, 2020c, art. 6 (h))."

This thesis addresses the question of how platform providers can influence design decisions regarding modules developed by other actors and the consequent effects on the applicability of design-oriented approaches to computer ethics. Furthermore, it also discusses the effect of regulatory measures on the applicability of these approaches to computer ethics. The respective findings provide core insights for subsequent reflections about the relationship between design- and policy-oriented computer ethics.

2.2.4 Artificial Intelligence Systems

In recent years, AI systems have been adopted in a wide variety of application contexts. As a result, AI has attracted considerable public, academic, and political attention. This is also true for computer ethics, which is increasingly concerned with ethical issues related to the use and design of AI systems. These include, for instance, unfair outcomes of decisions made by AI systems, the reliability of AI systems, the explainability of AI systems' decision-making processes, and the transformative effects of AI systems in society (Morley et al., 2020).

However, as is the case with the term 'platform,' definitions of 'Artificial Intelligence' vary widely. Because this thesis discusses AI systems related to, among other things, regulatory questions, especially concerning the proposal of the AI Act by the European Commission (European Commission, 2021c), it follows these regulatory proposals in their definition of AI. The AI Act applies a very broad definition of the term, taking into account various different techniques and approaches to AI. According to the AI

[16] Note that the term "gatekeeper" is defined more narrowly by the European Commission than the term "platform provider" that is used in this paper (see European Commission, 2020c).

Act, these include “a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning; b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems; and c) Statistical approaches, Bayesian estimation, search and optimization methods” (European Commission, 2021a, p. 1). As Hoffmann (2021) notes, such wide definitions covering various techniques and approaches are necessary from a regulatory point of view to conceptualize AI in a comprehensive but legally manageable manner.

The main research interest of this thesis in relation to AI systems is how different approaches to computer ethics can be applied in the context of AI systems. As with the other socio-technical systems discussed in this thesis, the high degree of distributed agency and power in the development of these systems is a major challenge in applying some approaches.

Technical Background:

The following paragraphs present each of the abovementioned approaches to AI. Morover, they outline the differences between them.

Machine-learning-based approaches to AI involve training a decision model on a large dataset, enabling the AI system to make predictions or decisions based on this decision model. Thus, the decision models in ML-based AI systems are not explicitly coded by their developers but are derived by identifying patterns in the training data (Russell & Norvig, 1995; Tsamados et al., 2022). This is done using algorithms that can automatically improve the model’s performance by adjusting its parameters based on the data it is exposed to. ML-based approaches to AI include supervised learning, unsupervised learning, and reinforcement learning (Bishop, 2006).

Statistical approaches to AI also involve using mathematical and statistical techniques to analyze data and make predictions or decisions. Unlike machine-learning-based approaches, however, these methods do not include training a decision model on data. Instead, they use existing statistical models and techniques to analyze the data and make predictions or decisions (Bishop, 2006; Datenethikkommission, 2019). Examples of statistical approaches to AI include regression analysis, which involves fitting a mathematical model to data to make predictions, and hypothesis testing, which involves using statistical methods to evaluate whether the data supports a given hypothesis.

Such data-driven approaches to AI need to be distinguished from knowledge-based ‘expert systems.’ In such systems, problem-specific knowledge of experts is formalized, allowing to automate rule-based decisions “on narrowly defined tasks” (Russell & Norvig, 1995, p. 255).

Even though the AI Act also covers rule-based expert systems, the vast majority of the obligations specified in the AI Act relate to requirements on components specific to systems relying on ML-based or statistical approaches to AI. Similarly, many discussions on AI ethics focus on ML-based systems specifically. As discussed more in-depth later in this chapter, this also holds true for this thesis.

Referencing Krafft et al. (2020, p. 121), Publication 3 explains the basic functioning of ML-based AI systems as follows:

Publication 3

> [...] especially algorithms of two types are involved: one algorithm that infers ‘decision rules from data’ and another one that ‘merely uses these decision rules to score or classify cases.’ The algorithm of the first type, ‘the learning method,’ and ‘the decision rules generated from it’ constitute the core of such ADM [17] systems, whereas the ‘scoring or classification algorithm, in contrast, is usually rather simple as it merely applies the trained statistical model.’

As mentioned earlier, a wide range of approaches to ML exist. However, while some of these approaches to machine learning raise method-specific ethical concerns, most of the AI ethics debate discusses AI systems on a more general level (cf. Jobin et al., 2019; Mittelstadt et al., 2016; Tsamados et al., 2022). This broader perspective focuses on the utilization of “vast data sets to train and feed machine learning algorithms that rely

[17] Krafft et al. (2020) avoid the term ‘Artificial Intelligence’. They refer to systems that “use information about entities and their behaviors in order to assign them a single numeric value by means of clearly defined instructions, that is, through an algorithm [...] that informs some decision or intervention that is either fully automated or occurs with a human in the loop” as Algorithmic Decision Making (ADM) system.

upon feedback loops to improve their own performance" (Yeung, 2019, p. 21). This thesis also adopts this perspective on AI systems.

Involved Actors:

The constellations of actors involved in ML-based AI systems are very heterogeneous. In the case of some systems, one actor may control all relevant tasks in the development of an AI system. These tasks include collecting or generating training data, data management and preparation, model development, and developing the application in which the model is embedded. In many other cases, however, these tasks are performed by different actors in a wide variety of constellations.

Publication 3 describes such interaction between actors with the illustrative example of an AI system designed to identify risk factors in patients' health records. The fictitious AI system detects patterns learned from patient data. Publication 3 argues that for such a system to work, several tasks need to be considered, which are potentially performed by various different actors:

Publication 3

> "[...] developing the underlying [Natural Language Processing (NLP)] capabilities, providing and preparing health records as training data, building a model that recognizes patterns based on this data, and using the system to classify or score unknown health records."

As Publication 3 elaborates, all these tasks may be performed by different actors.

Overall, the constellations of actors involved in AI systems are very heterogeneous and cannot be adequately described in a uniform manner. Accordingly, it is not possible to discuss how decision-making power is distributed among the actors involved in the development of AI systems in general terms. The thesis, therefore, introduces more concrete examples in the discussion of AI systems than is the case with the other socio-technical systems discussed in this thesis.

Research Interest:

As reflected in recently published AI ethics guidelines, academic articles, and regulatory proposals, AI systems pose numerous ethical issues (see, e.g., Fjeld et al., 2020; Jobin et al., 2019; Morley et al., 2020). While this thesis addresses many of these issues, the main focus is on the applicability of approaches to design- and policy-oriented computer ethics in the context of AI systems.

Publication 3 and Publication 4 address regulatory challenges that arise due to the distribution of decision-making power among the actors involved in the design of AI systems. These include the question of to which of the actors involved in the development processes of AI system policymakers meaningfully assign legal obligations. Here, Publication 3 focuses particularly on two frameworks proposed by the European Commission (European Commission, 2020a, 2021c). Publication 4 discusses the potential role of computer ethics in answering this question.

Publication 4 additionally takes challenges for design-oriented approaches to computer ethics into account. It identifies challenges that distributed agency and decision-making power pose for approaches such as VSD. Furthermore, it outlines opportunities for these approaches to either take advantage of or impact how power is distributed among the involved actors.

2.2.5 Similarities and Differences

As shown in the previous chapters, significant differences exist among the three socio-technical systems discussed in this thesis. However, there are also similarities between them. The following paragraphs discuss first the similarities and then the differences.

Similarities:

There are three main points of similarity. First, because they are considered to have a significant social impact, all three types of systems currently receive significant attention in computer ethics.

Secondly, all three types of systems are more complex socio-technical systems than the smaller and less complex ones discussed in most case studies of approaches such as VSD (see overviews in Friedman & Hendry, 2019; Winkler & Spiekermann, 2018). They have in common that power and agency are often distributed among the actors involved in ways that make the application of some approaches to design- and policy-oriented computer ethics particularly challenging. Because of the first similarity, it seems particularly appropriate for computer ethics to turn its attention to these systems. However, the second similarity raises questions about the applicability of some approaches to computer ethics.

Thirdly, all three types of systems are currently getting major attention from regulatory authorities. In the European Union, there is, for instance, the proposal for the Markets in Crypto-Assets Regulation (MiCA) (European Commission, 2020d) concerning blockchain-based cryptocurrencies, the Digital Markets Act (European Commission, 2020c) as well as the Digital Services Act (European Commission, 2020b) concerning platform ecosystems, and the proposal for the AI Act (European Commission, 2021c) concerning AI systems. All these regulations and regulatory proposals raise questions regarding the role of computer ethics in increasingly regulated environments.

Differences:

Yet, while the three types of systems pose some of the same challenges to computer ethics, they have significant differences. They serve different purposes, have different technical bases, different constellations of actors involved, and raise different ethical issues. The following paragraphs lay out differences that are important to answering the thesis' research question.

First, focusing on challenges to applying VSD, Publication 2 highlights differences that arise due to the different levels of decentralization of the systems:

Publication 2

> "Juxtaposing platform-based ecosystems and blockchain-based ecosystems suggests that considering the *level of decentralisation* of the ecosystem is of paramount importance for determining the kind of issues that might occur when accounting for power in the application of VSD. As Shilton and Greene (2019) demonstrate, actors like platform providers at the centre of more centralised ecosystems can assert their ideas of conceptions, weighings, and operationalisations of values to a large extent. [...]
> Conversely, in organisationally more decentralised ecosystems, there is, by definition, no central actor who can similarly assert itself. As shown in the case of open and permissionless blockchains, many actors have the power to impact decisions in the context of how human values are conceptualised, weighed, and operationalised in the design processes. The distribution of power in such ecosystems makes some form of deliberation and coordination inevitable to avoid gridlocks."

In the case of AI systems, the degree of decentralization cannot be generalized. As described in Chapter 2.2.4, the constellation of actors involved in AI systems is very heterogeneous. In the case of some systems, one actor can be in control of most relevant resources of an AI system. However, in many cases, these resources and components are developed, managed, and operated by different actors. Accordingly, the relationships among actors involved in AI systems cannot be described as schematically as those in platform ecosystems or blockchain-based systems. The level of decentralization of AI systems and its effect on the applicability of approaches to computer ethics, therefore, can only be determined on a case-by-case basis.

Secondly, the degree to which preexisting theoretical frameworks exist for analyzing and conceptualizing power in the respective systems differs. In this thesis, the concept of boundary resources served as a valuable theoretical lens through which platform ecosystems could be examined. Publication 2 explains the concept as follows:

Publication 2

> "Boundary resources are socio-technical manifestations of the platform provider's power to influence a platform ecosystem (Ghazawneh & Henfridsson, 2013), such as application programming interfaces (API), software development kits (SDK), legal guidelines, and application approval processes (Eaton et al., 2015; Karhu et al., 2018). As control points for a platform provider, boundary resources facilitate an arm's length relationship between the platform provider and service providers (Ghazawneh & Henfridsson, 2013). They offer the providers of peripheral applications access to a platform's resources while allowing the platform provider to retain influence over the platform (Eaton et al., 2015). Using boundary resources, a platform provider orchestrates its platform ecosystems and enables service providers to participate in and contribute to the platform's development (Eaton et al., 2015). Designing and implementing boundary resources is a balancing act of retaining power while supporting service providers to create independent platform-based innovation (Eaton et al., 2015)."

The respective literature reviews could not identify a counterpart to this concept for blockchain-based systems and AI systems.[18] However, while the concept is specific to platform ecosystems, analogous considerations on manifestations of power can also be made in the context of other systems. Thus, the thesis also analyzed how power manifests in the technical, social, political, and economic features of blockchain-based systems and AI systems and how involved actors can shape these features to gain or retain power within the respective system.

[18] The application of possible alternatives to the boundary resources concept is discussed in more depth in Chapter 6.

3 Methodology

The main methodological approach of this thesis is grounded in computer ethics. Chapter 3.1 provides background information on the general approach and the underlying perspective on computer ethics. As the thesis' research approach is mainly literature-based, the chapter starts by providing insights into the literature review process. Subsequently, the chapter describes the scope of computer ethics-based reasoning in this thesis, the approach to integrating theory and empirical observations, and the approach to integrating research perspectives of philosophy and computer science. Chapter 3.2 describes the specific research process for each of the publications.

3.1 Methodological Background

3.1.1 Literature Review

The thesis is based primarily on a critical examination and discussion of scientific literature from the fields of philosophy, computer science, and related disciplines. Additionally, gray literature such as ethics guidelines, (policy-related) whitepapers, policy proposals, comments and opinions on policy proposals, media articles, and technical documentation provided an essential resource for the research in this thesis.

The literature reviews provided the starting point for the thesis and the initial body of knowledge (cf. Snyder, 2019). More specifically, literature reviews concerning methods and concepts of computer ethics provided the theoretical basis of the research. In contrast, literature reviews concerning the features of socio-technical systems discussed in this thesis provided a basic understanding of the objects of research. This included literature with a technical focus as well as literature that discusses the societal impact of technical systems and related normative evaluations.

For the literature reviews, systematic database queries in scientific outlets were combined with the snowballing method, that is, "using the reference list of a paper or the citations to the paper to identify additional papers" (Wohlin, 2014). Continuous literature reviews were conducted to supplement the body of knowledge with more recent literature regarding the scientific and gray literature. This was particularly important in the case of literature on policy proposals, which were subject to regular renewal and modification.

3.1.2 The Scope of Computer Ethics-based Reasoning in the Thesis

Chapter 2.1 introduced the theoretical foundations of computer ethics. This chapter situates the thesis' research approach within the wider discipline. It sets out which perspectives of computer ethics the thesis adopts.

On the one hand, the thesis follows Moor (1985) in focusing on addressing policy vacuums. As Brey (2000, p. 10) argues, such policy-oriented computer ethics "takes as its point of departure a particular model of applied ethics that [...] aim[s] to clarify and evaluate morally controversial practices through an application and defense of moral principles." Yet, as outlined in more detail in Chapter 2.1.2, "mechanically applying legal and moral principles" is not always possible in computer ethics. This is because computer technology challenges existing concepts, and novel features of technology may make it "difficult to draw on traditional moral concepts and norms" (Johnson, 1999).[19] Therefore, addressing "conceptual muddle[s]" (Moor, 1985, p. 266) – that is, challenges for existing concepts caused by technological developments – is also an essential part of this thesis.

On the other hand, the thesis engages with design-oriented computer ethics. As outlined in Chapter 2.1.3, design-oriented computer ethics can take two forms. One type aims at "disclosing" (Brey, 2000, 2010) or "retrospectively analyzing" (Friedman et al., 2008; Friedman & Hendry, 2019) the ethical relevance of design features. The other, more constructive, type aims to proactively account for ethical values in technical design (Friedman et al., 2008; Friedman & Hendry, 2019; Nissenbaum, 2005). This thesis applies design-oriented computer ethics of the first type. The main focus of the design-oriented approach to computer ethics in this thesis is on analyzing and disclosing how features of the socio-technical systems in question – blockchain-based systems, platform ecosystems, and AI systems – affect the *applicability* of various approaches to computer ethics, including constructive design-oriented ones.

[19] See examples discussed in Chapter 2.1.2.

Thus, rather than focusing on inscribing specific values into the design of technology, it assesses conditions for successfully applying approaches to computer ethics.

Within computer ethics (and ethics more generally), such a self-reflective approach is not uncommon. In addition to outwardly directed interests concerning computer technology and related practices, computer ethics also has an inwardly directed self-reflexive interest. For instance, in recent years, VSD case studies have not only applied VSD but also have consistently identified limitations of the approach that arose in practice and thereby contributed to refining the approach (Friedman et al., 2021; Winkler & Spiekermann, 2018). Moreover, there are self-reflexive efforts that critically assess the general direction of entire research strands in computer ethics. Currently, for instance, there are major debates on the role that AI ethics can play in solving ethical issues of AI systems. Here, AI ethicists themselves reflect critically upon the field itself. Recent articles raised the question if AI ethics is doing more than providing "an easy alternative to government regulation" (Wagner, 2019), if it is primarily used for 'ethics washing' (Bietti, 2021; van Maanen, 2022), and if many of recently developed AI ethics guidelines are "useless," as they largely consist of "meaningless," "isolated," and "toothless" principles (Munn, 2022).

The discussions of the applicability of some approaches to computer ethics in blockchain-based systems, platform ecosystems, and AI systems in this thesis are ethical considerations in this self-reflexive mode. They aim to contribute to the critical literature on computer ethics and AI ethics.

3.1.3 Integration of Theory and Empirical Observations

Normative reasoning in computer ethics can be divided into two branches. First, there are theory-driven approaches. These approaches apply existing (normative) ethical theories to novel computer technologies and related practices. Secondly, there are pre-theoretical, interpretative approaches. These approaches take computer technology and related practices as a starting point for the analysis (Brey, 2000). The latter start with "common-sense definition[s]" (Brey, 2000, p. 12) of concepts such as moral values and refine them continuously during the research process. This is done through a co-development and reciprocal refinement of theory, on the one hand, and empirical observations of computer systems and related practices, on the other hand.[20]

Refraining from settling on ethical theories from the outset has significant advantages. First, in theory-driven approaches, the acceptance of an evaluation of the moral aspects of computer technology or related practices depends "on the acceptance of [that] particular moral theory" (Brey, 2010, p. 12). For instance, if a blockchain-based system is evaluated concerning if it makes trust in institutions obsolete, and a specific definition of trust is applied *ex ante*, the whole assessment might be rejected by readers who do not share the respective understanding of trust. As concepts like trust (fairness, bias, privacy, ...) are heavily contested, this poses a major problem for evaluating such technical systems. Starting with the technology and then interpreting it according to various perspectives on the concept of trust (fairness, bias, privacy, ...) avoids this issue (cf. Publication 1).

Secondly, as Moor (1985, p. 266) notes, "the mechanical application of an ethical theory" is often not possible in computer ethics because the transformative effects of "computerization" concern fundamental concepts of these theories. For instance, traditional concepts of property and ownership, which were conceptualized with scarce, physical goods in mind, do not readily apply to digital, non-rivalrous goods (see, e.g., Locke, 1988; Nozick, 2001; Rawls, 1999; Rousseau & Cole, 1992). Attempting to discuss property-related issues concerning digital goods based on a conceptual framework developed for physical goods – and thus based on the assumption that goods are scarce – is not a promising endeavor. Instead, discussing such issues requires considering the transformative effect of computer technology and reassessing and refining existing theories of property and ownership.

Thirdly, theory-driven approaches employ vocabulary and concepts from the start that may include empirical presuppositions in the analysis of observable phenomena. Therefore, observations can partly be based on preconceptions that are grounded in the underlying theory and its terminology (Brey, 2010, pp. 12–13). On the contrary, if theory and empirical observations continuously inform and (re-)shape each other, such blind spots are avoided.

[20] Thus, pretheoretical approaches to computer ethics do apply and develop theory. They just do not use theory as the starting point for the research.

Therefore, the research approach in this thesis follows this second approach. In the early phase, it is largely empirically guided. Based on initial findings, theories are introduced and refined to fit the application context. Subsequently, theory and empirical observations co-develop by informing and (re-)shaping each other. Chapter 3.2 explains this procedure in detail for each publication.

3.1.4 Integration of Philosophy and Computer Science

Given the object of observation – computer systems and related practices – the empirical observations partly require "considerable knowledge of the technological aspects of the system or practice" (Brey, 2000, p. 15) in question, and thus the integration of philosophy and computer science. Depending on the perspective on and approach to computer ethics, this integration can have different characters and take different forms.

For instance, VSD is based on an *iterative approach* in which more philosophical conceptual investigations, more social science-based empirical investigations, and more computer science and engineering-based technical investigations "inform and shape and reshape each other" (Friedman & Hendry, 2019, p. 35). The latter concern, for instance, how ethical values can be operationalized or which technical features of a system "enable, hinder, or even foreclose certain kinds of designs for supporting human activity" (Friedman & Hendry, 2019, p. 34). VSD prescribes neither a starting phase nor an order of phases. VSD practitioners can start with a value in conceptual investigations, a technology in technical investigations, or other points of departure and continue with iterations as deemed most productive.

In contrast, Disclosive Computer Ethics operates on three *sequential levels*: the disclosure level, the theoretical level, and the application level. The disclosure level is the initial level for Disclosive Computer Ethics. Here, a "computer system or software is analyzed from the point of view of a relevant moral value" (Brey, 2000, p. 15). Based on this analysis, on the theoretical level, "moral theory is developed and refined" (Brey, 2000, p. 15) to account for "the IT case at hand" (van den Hoven, 2008, p. 63). Lastly, on the application level, moral deliberation takes the form of joint consideration of moral theory, moral judgements or intuitions, and background facts or theories (Brey, 2010, p. 52). Especially on the first and the third level, integrating relevant domain knowledge from computer science is crucial to take into account the technological aspects of the system or practice under study.

In policy-oriented computer ethics, the integration does not follow a uniform pattern (as described for methods such as VSD and Disclosive Computer Ethics). Nevertheless, describing and analyzing computer systems is also necessary for policy-oriented computer ethics. Technical expertise regarding the to-be-regulated systems is relevant for drafting effective and ethically sound policy. For instance, understanding the functioning of AI systems and the constellations of actors involved in developing, managing, and operating these systems is necessary to identify appropriate addressees for specific obligations. Thus, policy-oriented computer ethics addressing such issues also requires integrating relevant domain knowledge from computer science. In this vein, for example, the European Commission's *High-Level Expert Group on Artificial Intelligence* has published a separate document on its understanding of the technical underpinnings of AI (HLEG-AI, 2018) as preparatory work for its *Ethics Guidelines for Trustworthy AI* (HLEG-AI, 2019).

As described in Chapter 3.1.1, the thesis' main source of technical expertise is literature from computer science. In addition to the approach of the literature review, the direct provision of literature on the respective systems by project partners from the IGT and GOAL projects contributed significantly to the research conducted in this thesis. Based on my interdisciplinary background, which is based on studying philosophy (M.A.) and computer science and philosophy (B.A.), I was able to integrate the technical and philosophical perspectives provided by this literature to a great extent autonomously.

However, as Brey (2000) argues, in some cases, an interdisciplinary background of a researcher can be insufficient to address specific research questions. In such cases, research should be conducted as a "cooperative venture" between computer scientists and philosophers. The direct and close cooperation with the respective project partners with computer science and information systems backgrounds was essential for refining the technical understanding of the assessed systems in this thesis. In particular, the project work in IGT contributed significantly to the thesis' perspective on platform ecosystems. In contrast, the project work in GOAL contributed considerably to the thesis' perspective on AI systems.

Furthermore, the research for Publication 2 was conducted as a cooperative venture with project partners of the IGT project. The research question was developed and addressed in close cooperation with my co-authors Christian Kurtz, Judith Simon, and Tilo Böhmann. Christian Kurtz and Tilo Böhmann particularly contributed their expertise on platform ecosystems. Conversely, Judith Simon and I provided a philosophical perspective.[21]

3.2 The Thesis' Approach in Practice

The following chapters describe the thesis' approach in practice. In particular, they describe how the integration of theory and empirical observations, as well as philosophy and computer science, occurred in detail. The chapter outlines the approach for each publication individually.

3.2.1 Publication 1: How Implicit Assumptions on the Nature of Trust Shape the Understanding of the Blockchain Technology

The starting point for the first publication, *How Implicit Assumptions on the Nature of Trust Shape the Understanding of the Blockchain Technology*, was the IGT project. In the early project phase, the blockchain technology was identified as a technology with novel and societally relevant governance features, thus fitting the project's scope. Following the empirically guided approach, Publication 1 places a first research focus on the technology, its underlying governance mechanisms, and discourses relating to these aspects.

In the initial literature analysis, trust emerged as a central concept used in the academic discourse to explain the technology's unique features. As outlined in Publication 1, the notion of trust appeared to determine many evaluations of the technology's features and capabilities:

[21] In Chapter 4, the contributions of the authors are described in more detail.

Publication 1

> “The blockchain technology is said to enable (Underwood, 2016, p. 16) and establish (Krishna, 2015) trust as well as to redirect it (Werbach, 2018, p. 30), to substitute for it (Freeman et al., 2020, p. 69), and to make it obsolete (Nakamoto, 2008, p. 8). Furthermore, there is disagreement on whom or what users have to trust when using the blockchain technology: (only) code, math, algorithms, and machines (Maurer et al., 2013; Nakamoto, 2008), or still (also) human actors (Botsman, 2017; Walch, 2019b; Werbach, 2018).”

Based on this identification of such vastly different perspectives, it became apparent that the authors understood trust to mean different things. However, as they did not always make their understanding of the nature of trust explicit, Publication 1 aims to disclose the implicit understanding of trust in the respective explanations of its alleged role in blockchain-based systems. It shows that some arguments rely on game-theoretical notions of trust, whereas others rely on motivation-attributing notions of trust. Both notions are discussed in-depth in Publication 1. In doing so, it provides a more profound understanding of the theoretical foundations and implications of the respective positions in the discourse.

These theoretical findings, in turn, inform an in-depth analysis of what perspective on blockchain-based systems various existing positions in the discourse take: a narrow technical perspective or a perspective that takes the broader socio-technical environment in which these systems are embedded into account.

Publication 1 shows that whether trust can be appropriately attributed to blockchain-based systems depends on the applied notion of trust *and* the perspective on blockchain-based systems. Based on a motivation-attributing account of trust, trust in blockchain-based systems is only conceivable if blockchain-based systems are understood as socio-technical systems that include human actors. In contrast, based on game-theoretical accounts of trust (which equate trust with reliability), trust in blockchain-based systems is also conceivable if blockchain-based systems are understood in a narrow technical sense. Based on these considerations, Publication 1 identifies implicit assumptions on the nature of trust and perspectives on blockchain-based systems in different positions in the discourse.

As the goal of the publication is to provide conceptual clarifications in the discourse on what role trust plays in blockchain-based systems, it refrains from making the case that one description of the role of trust in blockchain-based systems is correct and others are wrong. Its reasoning is rather analytical and descriptive than normative. It provides further theoretical grounding rooted in the philosophy of trust for various preexisting positions within the broader discourse on the role of trust in blockchain-based systems. Furthermore, it shows that views on the role of trust in blockchain-based systems, which at face value appear to be contradictory, are often based on 1) different understandings of the nature of trust and 2) different perspectives on what they consider to be part of a blockchain-based system.

3.2.2 Publication 2: Value Sensitive Design and power in socio-technical ecosystems

The research underlying Publication 2, *Value Sensitive Design and power in socio-technical ecosystems*, was conducted as part of the IGT project. It builds on previous research in the project, such as Publication 1 on the blockchain technology and further research focusing on platform ecosystems (see, e.g., Kurtz et al., 2018). Regarding both types of socio-technical systems, previous research in the project had addressed the means by which some actors involved in the respective system could interfere with the actions and (design-)decisions of other involved actors. In particular, platform providers have been identified as being able to use several tools at their disposal (APIs, app store approval processes, …) to influence the development of applications that can be offered on their platform. Because inscribing ethical values in technical systems requires a certain scope for design on the part of the developers, this poses challenges for approaches such as VSD or Values in Design.

Such challenges have also been raised in other research reflecting on the limitations of VSD. Some of these research articles introduce the concept of power as a theoretical lens. For instance, Friedman et al. (2021, p. 8) make the case that "identifying and potentially addressing power relationships" is a core challenge for the VSD approach, which has not been addressed sufficiently in the approach yet. They argue that most existing VSD studies "have not directly addressed the issue of power" (Friedman et al., 2021, p. 7). However, most VSD case studies concern rather stand-alone, monolithic applications in whose design agency is less distributed, and power-related issues are less prevalent (cf. Friedman & Hendry, 2019; Winkler & Spiekermann, 2018).

Therefore, power-related issues can be sidelined more easily in these case studies. In contrast, the power imbalances among actors involved in platform ecosystems and blockchain-based systems appear to question the applicability of approaches such as VSD more fundamentally.

Publication 2 attempts to address how the distribution of power within complex socio-technical systems, such as platform ecosystems and blockchain-based systems, poses a challenge for the application of VSD. As a starting point, Publication 2 defines power in a way that fits the scope of the research while still being open enough to not include empirical presuppositions in the analysis that might lead to missing relevant observable phenomena:

Publication 2

> "In this paper, the term 'power' refers to 'the ability of agents [...] to realize a certain outcome' (Brey, 2008, p. 75) – specifically design decisions – 'even against resistance' (Weber, 2019, p. 134). [...] In order to account for all the domains in which power can manifest itself, the paper adopts a systemic view on power, which 'regards power as the property of broader social, economic, cultural, and political networks, institutions, and structures' (Sattarov, 2019, p. 20) and focuses on how 'systems confer differentials of dispositional power on agents, thus structuring possibilities for action' (Haugaard, 2010, p. 425; see also Sattarov, 2019)."

Based on this understanding of power, a literature review was conducted to identify incidents in which actors involved in either blockchain-based systems or platform ecosystems used their power to "unilaterally intervene in design decisions that concern the realisation of a specific value" (Publication 2). Here, both empirical scientific studies and media articles were taken into account.

Further theoretical concepts relating to power in socio-technical systems, such as boundary resources (Eaton et al., 2015; Ghazawneh & Henfridsson, 2013; Karhu et al., 2018) and value levers (Shilton, 2012; Shilton & Greene, 2019), were applied to refine the theoretical lens for analyzing and interpreting the results. Based on the analysis of the incidents, Publication 2 identifies four factors that determine the effects that the distribution of power in different socio-technical systems can have on the applicability of VSD: 1) the level of decentralization of the ecosystem, 2) if VSD is applied at the core or periphery, 3) when power can be exercised (temporality), 4) and the phase of VSD (conceptual, empirical, and technical) that power can be exercised in.

Subsequently, Publication 2 highlights concrete challenges for applying VSD in the discussed socio-technical systems and outlines ways to address these challenges. Furthermore, it emphasizes opportunities for applying VSD in these systems.

Regarding ways to address the challenges for VSD, Publication 2 discusses a further type of empirical material: current policy proposals that attempt to limit the ability of platform providers to restrict the scope for design of developers of applications offered on their platform. It argues that establishing new oversight institutions and further regulatory measures can support dealing with powerful actors in socio-technical systems. The proposal for the Digital Markets Act serves as a prime example of how the quasi-monopolistic standing of many platform providers can be curbed. By making this argument, Publication 2 establishes a link between design- and policy-oriented computer ethics. This link becomes a central theme of Publication 4 and the thesis at large.

3.2.3 Publication 3: Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed by the European Commission

The GOAL project was the starting point for Publication 3, *Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed by the European Commission*. Setting a clear policy-oriented focus, the project aimed to develop and formulate potential strategies for the regulation of and by algorithms. In the project, there was a particular focus on AI systems, which Publication 3 adopted accordingly. The contribution of the project team from the University of Hamburg was particularly concerned with regulation that aims to uphold ethical values and principles. These included, for instance, fairness and solidarity (cf. Publication 3; Rudschies, 2023).

Previous work by project partners and (at the time) recent whitepapers by the European Commission (cf. European Commission, 2019, 2020a) already had identified distributed agency in development processes as a challenge for regulatory proposals. Publication 3 contributes to the development of policies in the face of this challenge. In particular, the research underlying Publication 3 focused on the question of how appropriate addressees for obligations in AI regulation can be identified and selected.

Initially, the research question was formulated as an open question. However, during the research process, the European Commission published its proposal for the AI Act. To account for this development, Publication 3 now discusses two frameworks for assigning obligations comparatively: the framework outlined in the proposal for the AI Act and the framework outlined in the European Commission's whitepaper *On Artificial Intelligence – A European approach to excellence and trust*, which preceded the AI Act.

In the first step, an in-depth analysis of the respective whitepapers and policy proposals was conducted. This process revealed the two proposal's respective frameworks for assigning obligations in AI regulation:

Publication 3

> "The EC's whitepaper *On Artificial Intelligence* proposes a capability-based approach for assigning obligations arguing that 'the actor(s) who is (are) best placed to address' the respective issue should be obliged to do so (European Commission, 2020a). On the contrary, the AI Act argues that the 'majority of all obligations' should fall on the person or body 'placing [the AI system] on the market or putting it into service under its own name or trademark' (Veale & Zuiderveen Borgesius, 2021) and thus focuses on rather fixed addressees."

In the next step, the two proposals were evaluated as to whether or not the frameworks take into account the complexity and heterogeneity of the constellation of actors involved in the development of AI systems and where in the development process ethical issues arise. For this purpose, it was necessary to integrate perspectives of computer ethics on AI systems with technical expertise on AI systems.

In order to answer the questions, a comprehensive literature review was carried out to reveal 1) what components AI systems consist of, 2) which actors are involved in the development of these components, 3) whether the resulting actor constellations are rather homogeneous or heterogeneous, 4) in which tasks of the process of developing, deploying, and operating AI systems ethical issues can arise, and 5) which of the involved actors can address these issues. The literature review included resources from both computer science and AI ethics. Furthermore, it considered (at the time) recent gray literature reporting on, e.g., AI business models (e.g., Keller et al., 2018) to complement the findings of the literature review and account for persisting gaps in the body of knowledge.

Drawing on this literature analysis, the paper maps the ethical issues of AI systems to the various tasks in the development process of AI systems. Thereby, Publication 3 links the technical expertise on the design and development of AI systems to the philosophical views on the ethical issues of AI systems.

Based on these findings (manifesting in Section 2 and Section 4 of Publication 3), the two frameworks for assigning obligations in AI regulation were evaluated. Here, the main focus was on the question of to what extent the two frameworks account for how the various tasks in the process of developing, deploying, and operating AI systems are distributed among involved actors. The results of this analysis, in turn, inform a discussion of whether the two frameworks are appropriate to address ethical issues of AI systems.

Publication 3 uses these findings to judge the merit of the EC's shift from the framework outlined in the whitepaper *On Artificial Intelligence* to the framework outlined in the AI Act. Because the latter contains fewer ambiguities and ensures that there is always a well-defined and identifiable actor to whom obligations can be assigned, it concludes that the shift from a framework focused on capabilities to a framework focused on fixed addressees is appropriate.

3.2.4 Publication 4: Reexamining computer ethics in light of AI systems and AI regulation

Publication 4, *Reexamining computer ethics in light of AI systems and AI regulation*, builds heavily on the former publications and uses their results as a starting point. In particular, it reflects on two findings: 1) challenges that distributed power and agency in development processes pose for policy-oriented computer ethics identified in Publication 3 and 2) challenges that distributed power and agency pose for design-oriented approaches to computer ethics identified in Publication 2. The first finding was discussed in relation to two different policy proposals concerning AI systems by the European Commission. The second finding was discussed in relation to the application of VSD in the context of platform ecosystems and blockchain-based systems. In line with the self-reflexive approach to AI ethics described in Chapter 3.1.2, Publication 4 evaluates the significance of these findings for computer ethics more generally.

To conduct the analysis on a specific case, Publication 4 focuses on AI systems. This selection was made because there is sufficient literature on both design-oriented and policy-oriented computer ethics in the case of AI systems. As the findings of Publication 2 are based on an analysis of the applicability of VSD in platform ecosystems and blockchain-based systems, the extent to which the findings are transferable to AI systems had to be examined first.[22] Publication 3's analysis of the technical properties of AI systems and the constellation of actors usually involved in their development, management, and operation informed an adaptation of the findings of Publication 2 to the new application context. On this basis, some cases analogous to those discussed in Publication 2 could be identified for AI systems. These concerned, for instance, the impact of regulation and power imbalances among involved actors on the applicability of design-oriented computer ethics.

Moreover, to reexamine design-oriented and policy-oriented approaches to computer ethics on an equivalent level of abstraction, it was necessary to examine the findings of Publication 3 independent of the specific context of the frameworks for assigning obligations that Publication 3 discusses. To discuss the challenges identified in Publication 3 independent of this particular context, the concept of power was reintroduced:

Publication 4

> "This article adopts a definition of power focusing on outcomes, according to which power is the 'ability of agents' to 'realize a certain outcome' or 'bring about certain [...] state of affairs' (Brey, 2008; see also Dowding, 1996)."

Based on this groundwork, the challenges for design- and policy-oriented computer ethics in AI systems were examined alongside each other. As mentioned at the beginning of this chapter, this examination also revealed new opportunities for both design- and policy-oriented computer ethics. Publication 4 outlines two of these in

[22] See the similarities and differences of the discussed socio-technical systems outlined in Chapter 2.2.5.

particular. First, it shows that computer ethics can be used to shape how power manifests among the actors involved in an AI system. It argues that this allows enabling specific actors to apply approaches to computer ethics more effectively in the future. Secondly, by considering how the ability to influence design decisions (even against resistance) differs among actors involved in a socio-technical system, powerful individual actors can be identified and encouraged (or forced) to shape the respective socio-technical system in accordance with specific ethical values or principles.

Moreover, Publication 4 shows that policy- and design-oriented computer ethics can engage with the same actors, computer systems, and value conflicts. This suggests that the simultaneous application of design- and policy-oriented approaches to computer ethics (by different actors) could create synergies or conflicts. While synergies between the approaches had already been described in the theoretical literature on computer ethics (e.g., Brey, 2000), conflicts have not been covered in-depth. However, individual research articles with a more practical focus provided sufficient evidence for such conflicts. These are discussed as practical examples. Furthermore, it was possible to draw directly on examples from Publication 3 that demonstrate how design elements of AI systems impact policymaking. Publication 4 uses some of these examples to describe how conflicts could arise between the approaches in case they are applied with conflicting values in mind.

4 Publications

This thesis is based on four peer-reviewed research articles. The articles were produced and published continuously throughout the research process. All four articles have been published in academic journals, namely Philosophy & Technology (Springer), Internet Policy Review (Alexander von Humboldt Institute for Internet and Society), Digital Society (Springer), and AI & Ethics (Springer). All publications directly or indirectly contribute to the subject of this thesis and to answering its research questions. The following overview provides general information on each publication. Additionally, there is a title page attached to each publication, providing the most relevant information again.

Publication 1

Title:
How Implicit Assumptions on the Nature of Trust Shape the Understanding of the Blockchain Technology

Author(s):
Mattis Jacobs

Journal:
Philosophy & Technology; ISSN: 2210-5441 (electronic); 210-5433 (print)

Publisher:
Springer

Author Contributions:
Single-authored publication.

Chapter in the thesis:
7

Publication 2

Title:
Value Sensitive Design and power in socio-technical ecosystems

Author(s):
Mattis Jacobs, Christian Kurtz, Tilo Böhmann, Judith Simon

Journal:
Internet Policy Review; ISSN: 2197-6775

Publisher:
Alexander von Humboldt Institute for Internet and Society

Author Contributions:
Mattis Jacobs conceived the article, conducted the main part of the literature analysis, and wrote the first draft of all sections except those specifically related to platform ecosystems and information systems concepts. Christian Kurtz provided feedback on the conception of the article, conducted the literature review regarding platform ecosystems and information systems concepts, and wrote the first draft of the sections referring to these concepts. Moreover, he provided continuous feedback on all parts of the manuscript in the writing process. Judith Simon and Tilo Böhmann provided feedback on the conception of the article and commented on earlier versions of the manuscript. All authors read and approved the final version of the manuscript.

Chapter in the thesis:
8

Publication 3

Title:

Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed by the European Commission

Author(s):

Mattis Jacobs, Judith Simon

Journal:

Digital Society; ISSN: 2731-4669 (electronic); 2731-4650 (print)

Publisher:

Springer Nature

Author Contributions:

Mattis Jacobs conceived the article, conducted the literature analysis, and wrote the first draft of the manuscript. Judith Simon provided feedback on the conception of the article and commented on earlier versions of the manuscript. All authors read and approved the final version of the manuscript.

Chapter in the thesis:

9

Publication 4

Title:

Reexamining computer ethics in light of AI systems and AI regulation

Author(s):

Mattis Jacobs, Judith Simon

Journal:

AI & Ethics; ISSN: 2730-5961

Publisher:

Springer Nature

Author Contributions:

Mattis Jacobs conceived the article, conducted the literature analysis, and wrote the first draft of the manuscript. Judith Simon provided feedback on the conception of the article and commented on earlier versions of the manuscript. All authors read and approved the final version of the manuscript.

Chapter in the thesis

10

5 Contributions

Each of this thesis consists of four independent publications, each of which represents an independent research contribution. Moreover, the thesis also provides a distinctive overall contribution beyond the contributions of the individual publications. This overall contribution concerns a broader reexamination of different strands of computer ethics. Publication 4 – *Reexamining computer ethics in light of AI systems and AI regulation* – already engages in this reexamination, building on the results of Publication 2 – *Value Sensitive Design and power in socio-technical ecosystems* – and Publication 3 – *Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed by the European Commission*.

However, the scope of Publication 4 is limited to AI systems, whereas the thesis expands this perspective. Building on Publication 1 – *How Implicit Assumptions on the Nature of Trust Shape the Understanding of the Blockchain Technology* – as well as Publications 2 and 3, it shows that the reasoning of Publication 4 also applies to other socio-technical systems. Thus, the thesis as a whole has more generalizable implications for computer ethics than Publication 4.

The following chapters outline these implications. Chapter 5.1 presents the contributions to design-oriented computer ethics, while Chapter 5.2 presents the contributions to policy-oriented computer ethics. Finally, Chapter 5.3 outlines implications for the integration of design- and policy-oriented computer ethics.

5.1 Contributions to Design-oriented Computer Ethics

5.1.1 Identification of Novel Challenges

Publication 2 and Publication 4 show how technical, social, economic, and political features of more complex socio-technical systems can limit the ability of actors involved in these systems to realize certain outcomes. This also concerns the ability to design technical components of the system in accordance with ethical values and principles. Consequently, if applied in the context of such socio-technical systems, design-oriented computer ethics needs to consider the abilities of involved actors to shape components of the system to a larger extent. Furthermore, it needs to take into account what constrains the ability of various actors and which actors set these constraints. Focusing on AI systems, Publication 4 highlights the following challenges for design-oriented computer ethics:

Publication 4

> “For more analytical approaches to design-oriented computer ethics, such as Disclosive Computer Ethics, the question arises which of the involved actors has the ability to address problematic ethical features of computer systems that are integrated into larger socio-technical systems once these features have been disclosed. For more constructive approaches to design-oriented computer ethics, such as Value Sensitive Design, the question arises which actors involved in a socio-technical system can assert design decisions that align with specific ethical principles or values and can, therefore, successfully apply these approaches. Conversely, they also have to engage with the question of which actors lack the ability to apply them successfully and how they can change this circumstance.”

This line of reasoning does not only apply to AI systems. Accounting for power is a more general issue when applying design-oriented approaches to computer ethics in larger and more complex socio-technical systems.[23] For example, as described in Publication 2, platform providers can determine to what extent service providers and app developers can access data by using API design. Li et al. (2021) show that changes in the API design of platforms, specifically iOS and Android, impact how app developers account for privacy in the development of applications offered on these platforms. Furthermore, by making use of platform policies and app store approval processes, platform providers can impact the extent to which app developers and (other) service providers can design applications or provide services in accordance with specific values as they see fit (Li et al., 2021). These phenomena are already discussed in the computer science and information systems literature but have only started to receive attention in design-oriented computer ethics (see, e.g., Friedman et al., 2021; Shilton & Greene, 2019).

Moreover, policymakers increasingly influence the scope for design of developers. Many of the regulations drafted by policymakers aim to uphold ethical values and

[23] However, the details of how exactly power-related issues manifest depend on various features of socio-technical systems. Here, Publication 2 names, for instance, the level of decentralization.

principles such as fairness or transparency in AI systems. Yet, as a side-effect, these legally binding regulations potentially also limit the ability of developers, operators, and users of computer systems to negotiate and act on what they consider to be ethical behavior and design. For design-oriented computer ethics, such regulations, therefore, pose a challenge if they aim to uphold ethical values or principles that conflict with the values and principles upheld by developers. Publication 4 provides an example:

Publication 4

> "[...] privacy regulations can hamper bias mitigation strategies that require integrating more data (Jobin et al., 2019). Furthermore, values like fairness can be defined in various conflicting ways. Thus, requiring an AI system to make fair decisions according to one definition of fairness makes it impossible to achieve fair decisions according to a conflicting definition of fairness (Binns, 2018). Therefore, such regulatory interventions can obstruct or compel design decisions that promote or demote the realization of specific values (Jacobs et al., 2021). Consequently, they can reduce the developers' scope for design and hamper their ability to negotiate and account for values themselves or in accordance with further stakeholders."

Policy proposals directed at other socio-technical systems – such as the Digital Markets Act and the Digital Services Act (European Commission, 2020b, 2020c) which aim at platform ecosystems – indicate that similar circumstances will likely occur in these contexts.

Lastly, Publication 1 shows that not only accounting for power is difficult in large-scale socio-technical systems with complex actor constellations. It highlights that achieving an agreement on how values are defined and framed can be equally challenging. Yet, defining values is a major task in approaches such as VSD (Friedman et al., 2008; Friedman & Hendry, 2019) or Values in Design (Nissenbaum, 2005). This poses an additional challenge for applying approaches to design-oriented computer ethics in such contexts.

To give an example, when Vermaas et al. (2010, p. 497) elaborate on using VSD to account for trust in technical design processes, they characterize trust as "a distinctively moral phenomenon." According to them, "[t]rust between people is crucially concerned with assumptions or beliefs about the benevolence and moral motivation of others." Within their study, this perspective is not challenged, although competing perspectives are acknowledged. The same holds true for other VSD case studies focusing on trust with fewer actors involved in the design process (Friedman et al., 2000; Tauro, 2021).

Yet, as Publication 1 shows, many scholars, activists, developers, and commentators involved with blockchain-based systems reject such more demanding accounts of trust that go beyond rational expectations. Instead, they adhere to a rational expectation-based account of trust that equates trust with reliability.[24] Therefore, Publication 1 concludes that finding common ground on such fundamental questions among stakeholders would most likely be difficult. With an increasingly large number of actors involved, it becomes even more challenging to determine not only *which* values should be considered in design decisions but also *how* values are defined or best understood in the first place. So far, method overviews for the application in VSD (Friedman et al., 2017; Friedman & Hendry, 2019) lack tools for resolving conflicts that arise if there is not one actor ultimately in charge of the process.

5.1.2 Identification of Novel Opportunities

As Publication 4 demonstrates for AI systems, there are not only challenges arising for design-oriented computer ethics, but there are also new opportunities. These take the form of new options for action and fields of activity when applying approaches to design-oriented computer ethics. Publication 4 outlines two ways in which approaches to design-oriented computer ethics can take advantage of increasingly complex actor constellations in many socio-technical systems.

First, computer ethics can be used to shape how power manifests in the technical, social, political, and economic features of socio-technical systems. In this way, it can help actors to more easily apply policy- or design-oriented approaches to computer ethics in the future. This approach does not directly promote the consideration of specific ethical values or principles within the given system. Instead, it shapes the conditions for successful future applications of computer ethics. Secondly, computer ethics can be used to identify and encourage (or force) powerful actors to shape the

[24] Publication 1 explains these accounts of trust in detail.

respective socio-technical system in accordance with specific ethical values or principles.

Publication 2 and Publication 4 exemplify the first approach. Publication 4 shows how making an AI system's features more transparent allows for a broader conversation among the involved actors about the system's values (Slota, 2020). Thereby, this approach can foster the ability of actors to ensure that further design decisions align with the ethical values and principles they uphold.

Conversely, Publication 2 shows that the Bitcoin blockchain is designed to prevent any central actor from gaining exclusive power over the system. In other words, no individual actor is supposed to be unilaterally able to make changes to the system's design or make far-reaching decisions for its use, such as what transactions are considered valid or not (De Filippi & Wright, 2018; Nakamoto, 2008; Swan, 2015; Werbach, 2018). Accordingly, the design is such that a variety of involved actors can influence the extent to which ethical values and principles are upheld when the system is used or modified.

Publication 2 also exemplifies the second approach. It shows that computer ethics can be used to identify and encourage (or force) powerful actors in a socio-technical system to shape the system according to certain ethical values or principles. One example discussed in the paper illustrates how platform providers can influence the design of applications that are distributed on the platform by external actors. This can be achieved through explicit means such as app store approval processes.

Yet, Publication 2 demonstrates that platform providers can also use more subtle measures to exert influence. For instance, building on Ausloos and Veale (2020), the article shows that platform providers can use "strategic changes to an API" (cf. Ausloos & Veale, 2020) to either restrict access to data or provide privileged access. This, in turn, can potentially damage or even break the business models of service providers or app developers or, conversely, can establish a competitive advantage. Accordingly, developers of applications are highly incentivized to "adapt their design values to fit those of the platform on which they build" (Greene & Shilton, 2017, p. 16; see also Shilton & Greene, 2019).[25]

Accordingly, Publication 4 summarizes this new field of activity for design-oriented computer ethics as follows:

Publication 4

> "Thus, if design-oriented computer ethics is applied by actors in a dominant position in a socio-technical system, these actors can not only affect the design of technical components or applications which they are designing. To a varying degree, they can also shape the broader socio-technical system by co-determining if (and if so, how) values are accounted for in the system at large."

However, the two approaches can be in conflict with one another. Attempts by an actor to shape a socio-technical system in its entirety (i.e., including technical components and applications that are developed and operated by other actors) require that this actor has control over design decisions and other actors do not. If, for instance, platform providers in a platform ecosystem restrict access of service providers to certain types of data to enhance privacy for its users, service providers cannot negotiate tradeoffs between privacy and other involved values themselves. Consequently, there is a conflict between applying design-oriented computer ethics to *determine* design decisions based on ethical considerations or applying it to *enable* these actors to engage in ethical considerations in design processes themselves (cf. Publication 4).

5.2 Contributions to Policy-oriented Computer Ethics

5.2.1 Identification of Novel Challenges

In order to formulate effective policies for the ethical design and use of computer technology, policy-oriented computer ethics must consider the increasingly complex

[25] For this discussion, Publication 2 also introduces Shilton and Greene's concept of value levers. These allow platform providers to reward or punish external actors based on their performance in accounting for the values that the platform promotes. For further elaborations on the concept "value levers" see Publication 2 as well as Shilton (2012) and Greene and Shilton (2017).

constellations of actors involved in many socio-technical systems. Publication 3 and Publication 4, in particular, outline the challenges these complex actor constellations pose for formulating effective policies.

Publication 3 highlights the difficulties of assigning obligations in regulating socio-technical systems, specifically AI systems. These arise from the difficulties of determining appropriate addressees for obligations among the actors involved in developing and using such systems. The article evaluates two frameworks for assigning obligations in AI regulation. Both frameworks have been proposed by the European Commission (European Commission, 2020a, 2021c). The identified challenges include, for example, the lack of clarity about which actors are best placed to address the ethical issues of AI systems and the lack of information that many actors have to address these issues effectively.[26]

Based on these findings, Publication 4 reflects on implications for policy-oriented computer ethics. It argues that

Publication 4

> "[...] the rise of more vast and complex socio-technical systems such as AI systems forces policy-oriented computer ethics to determine not only what ethical practices relating to computer technology are (Moor, 1985) but also which actors have the ability to engage in these practices and which actors the respective obligations should be assigned to. To ensure the intended effects of policy measures, it is crucial to account for the involved actors' power to achieve specific outcomes."

Thus, if socio-technical systems become more complex, policy-oriented computer ethics needs to examine the ability of actors to achieve certain outcomes to a larger extent. If policymakers assign obligations to ensure or prohibit certain features of a socio-technical system to actors incapable of abiding by them, assigning these obligations will not achieve the desired results. Publication 4 discusses practical examples of this challenge and shows that while this observation may seem trivial in theory, it leads to major challenges in practice.

[26] For example, regarding features of a set of training data, see Gebru et al. (2021).

Publication 4

> “For instance, if policy-oriented computer ethics seeks to ensure that actors involved in AI systems warrant that potential bias in training data does not lead to biased decisions that harm individuals, it is challenging to determine which involved actors can or should be addressed: actors in charge of data collection and management (to ensure that there is no bias in the training data), actors in model development (to ensure that compensatory bias is applied so that decisions are unbiased), operators (to question decisions and not rely on them in cases that decisions might be biased), or providers.”

Thus, policy-oriented computer ethics must address the question of which obligations can and should be assigned to which actors. This is particularly difficult when constellations of actors are very heterogeneous in the type of system that a policy aims at. This, for instance, applies to AI systems.[27] Formulating policies that are sensitive to this heterogeneity but still sufficiently generalizable is a particularly challenging task.

As Publication 1 and Publication 2 show, similar challenges also arise in the context of blockchain-based systems. In particular, the design features aimed at achieving disintermediation and preserving privacy pose challenges for know-your-customer- and anti-money-laundering-regulation (De Filippi & Wright, 2018).[28]

5.2.2 Identification of Novel Opportunities

The way that policy-oriented computer ethics can approach power-related issues largely mirrors the way outlined in Chapter 5.1.2 for design-oriented computer ethics. First, it can consider how the ability to influence design decisions (even against resistance) differs among actors involved in a socio-technical system. This, in turn, makes it possible to advocate for policies that would oblige particularly powerful actors to shape the system in accordance with certain ethical values or principles. Secondly, it can influence how power manifests among these actors in order to shape their ability to achieve specific outcomes. In doing so, policy-oriented computer ethics can set the conditions for the success of future applications approaches to computer ethics.

[27] See Chapter 2.2.4.

[28] Chapter 5.3 discusses this issue in more depth.

As Publication 3 shows, the proposal for the AI Act is a prime example of a policy proposal that considers how the ability to influence design decisions differs among actors involved in a socio-technical system. It assigns most of the obligations to the providers and, to a lesser extent, to the users of AI systems (European Commission, 2021c; see also Veale & Zuiderveen Borgesius, 2021).[29] This contrasts with proposals of earlier whitepapers (see, e.g., European Commission, 2020a). In these whitepapers, the European Commission proposed that the obligations to ensure that certain properties of socio-technical systems are met should be assigned to those actors who are "best placed" to address them.

However, according to the AI Act, providers are expected to use their position to force other involved actors to ensure adequate data governance, provide technical documentation, or establish a quality management system (European Commission, 2021c, art. 10; art. 11; art. 17). Thus, by identifying powerful actors who are able to shape the system at large, the proposal for the AI Act avoids that policymakers need to evaluate the capabilities of individual actors themselves on a case-by-case basis.

Alternatively, policy-oriented computer ethics can advocate for policies that change how power manifests in a socio-technical system's features. This approach can be used to enable individual actors to account for ethical values in design processes more effectively. Publication 2 shows that this can be necessary because dominant actors within a socio-technical system can use the system's technical, social, political, and economic features to retain control over how ethical values and principles are accounted for in the system at large. Thereby, these actors can prevent others from taking ethical values and principles into account in the design process in the way they deem appropriate.

Publication 2 discusses the Digital Markets Act (European Commission, 2020c) as an example of how this approach can work. It shows how policies can mitigate power imbalances that hamper the applicability of approaches to computer ethics for some actors.

[29] While the AI Act's approach is a prime example of identifying powerful actors and obliging them to use their position to shape a system according to particular values, it is not without its own problems. The AI Act recognizes that in certain circumstances, providers may not be able to fulfill their obligations. This is the case, for example, when AI systems are used in contexts for which they are not intended. In such cases, the AI Act, therefore, places more obligations on users and fewer on providers of AI systems. Publication 3 discusses this issue in more detail.

The proposal for the Digital Markets Act encompasses various propositions that aim at preventing platform providers in platform-based ecosystems from exploiting their quasi-monopolistic position. For example, according to the European Commission's proposal, service providers must not be precluded from offering services on other platforms (at different prices). Furthermore, platform providers must offer data portability to end-users or service providers who want to switch platforms. According to Anderson and Mariniello (2021), such attempts aim to curb "winner-takes-all dynamics," which put quasi-monopolistic platform providers in an extremely powerful position in their respective ecosystems. Thus, these proposals seek to empower service providers and end-users vis-à-vis platform providers.[30] In turn, this gives more leeway to developers of peripheral applications in accounting for values in their applications' design.

Publication 2 argues developments such as the implementation of the Digital Markets Act

Publication 2

> "[...] would provide more choices to select a suitable platform (or suitable platforms) for developers of peripheral applications and make it a more viable option to integrate this selection process [of a platform that aligns with the developer's values] in the application of VSD."

Thus, just as design-oriented computer ethics, policy-oriented computer ethics can be applied not only to advocate for policies that encourage or force the consideration of ethical values by dominant actors in a socio-technical system. Policy-oriented computer ethics can also be used to advocate for policies that shape how power is distributed among the actors involved in a socio-technical system. Thereby, it can co-determine if and by whom approaches to computer ethics can be applied effectively.

5.3 Contributions to Integrating Design- and Policy-oriented Computer Ethics

The thesis also contributes to the understanding of how design- and policy-oriented computer ethics relate to one another. It challenges previous accounts of the

[30] Please note that there are so far no empirical assessments of the effectiveness of these approaches, as the AI Act and the Digital Markets Act are currently only regulatory proposals.

relationship between the two approaches to computer ethics and develops a more multifaceted understanding. For instance, Brey (2000) argues that design-oriented computer ethics (particularly Disclosive Computer Ethics) is complementary to "mainstream" or policy-oriented computer ethics. In contrast, the thesis contends that such assessments need to be re-evaluated. It argues that while the two approaches can be complementary, that is, create synergies, they can also be at odds with each other.

Focusing on AI systems, Publication 4 demonstrates that policy-oriented and design-oriented computer ethics can engage with the same actors, computer systems, value conflicts, and, more generally, state of affairs. Both design- and policy-oriented computer ethics can be applied to target specific actors involved in an AI system to encourage or force them to ensure that specific ethical values or principles are accounted for in the system at large. Furthermore, Publication 4 shows that both approaches can be applied to influence how an AI system's technical, social, political, and economic features co-determine how power manifests among the involved actors. In turn, this affects the extent to which these actors can successfully shape the system in accordance with their values and principles. Publication 4 elaborates on how this can lead to conflicts:

Publication 4

> "If design-oriented computer ethics is applied in contexts where policy constrains design decisions, developers and design-oriented computer ethicists need to take this circumstance into account. [...] [P]olicies can affect the application of design-oriented computer ethics in two ways. On the one hand, they can affect the consideration of specific values in design decisions. On the other hand, they can affect the ability of actors involved in a socio-technical system to influence design decisions and thus shape technology in line with ethical values and principles. This can lead to conflicts if either the respective approaches to computer ethics promote conflicting values (or operationalizations of values) or if one approach aims at enhancing the ability of specific actors to achieve their goals in a way that counteracts the other approach."

Publication 4 exemplifies such interdependencies with various cases related to AI systems. For instance, it shows that data protection regulations can conflict with design-oriented approaches to bias mitigation that rely on integrating more and more diverse data (see also Jobin et al., 2019).

Furthermore, there are examples in which design-oriented computer ethics is explicitly designed to avoid or counteract regulation. The Bitcoin blockchain is such an example. The creator[s] of Bitcoin developed the underlying blockchain technology "as a solution to the [perceived] problem of government oversight of value-based transactions" and to escape the so-called "prison of regulation" (Werbach, 2018, p. 158).

Bitcoin aims to achieve this through disintermediation. In the traditional financial system, intermediaries are simultaneously heavily regulated entities and access points for regulatory authorities to approach further actors. Such intermediaries (e.g., banks) enforce know-your-customer- and anti-money-laundering-regulation. Based on ethical assessments of problematic uses of cryptocurrencies, policy-oriented computer ethicists might suggest implementing similar policies also for systems like Bitcoin. However, these principles are harder to enforce in such systems than in the traditional financial market, as they circumvent these commonly regulated entities (Wright & De Filippi, 2015).[31]

These findings are consistent with recent studies highlighting the challenges of regulating blockchain-based systems due to their design features. For instance, De Filippi et al. (2022) describe the blockchain technology underlying Bitcoin as *"alegality by design,"* as the system's design intentionally situates its use "beyond the boundaries of existing legal orders."

In this line of thinking, the policy vacuum regarding the use of the technology – a typical problem of policy-oriented computer ethics (cf. Moor, 1985, p. 266) – is not a mere byproduct of the novel actions that the technology enables. It is the explicit goal of developing the technology in the first place. These insights further support arguments raised in Publication 4 concerning potential conflicts between design- and policy-oriented computer ethics and the resulting multifaceted nature of the relationship between these two approaches.

[31] Whether the approach hinders or prevents regulatory action is a matter of debate. For an in-depth analysis, see De Filippi and Wright (2018).

However, Publication 4 also acknowledges that policy- and design-oriented computer ethics can complement each other, as argued by Brey (2000) and others. For instance, as Publication 2 discusses in-depth, dominant actors in socio-technical systems can often prevent developers of technical components from accounting for ethical values and principles in design processes. This can happen, for example, if accounting for these values conflicts with the dominant actor's commercial interests. Yet, Publication 4 makes the case that adequate policies can reduce or dissolve such conflicts:

Publication 4

> "[…] policy-oriented computer ethics can promote regulation that establishes a threat of fines for not ensuring that technology design accounts for specific (operationalizations of) values. In doing so, it can change the cost-/benefit-analysis of these actors and soften or end the resistance to design decisions in accordance with specific ethical values."

Conversely, as Publication 3 shows, the design of technical components of socio-technical systems also co-determines how well the respective system can be regulated. For instance, designing AI systems more transparently helps to identify problematic uses and, in turn, the "formulation and justification of policies for the ethical use" (Moor, 1985, p. 266) of such systems. Thus, this thesis does not contradict Brey's perspective that design- and policy-oriented computer ethics complement each other. Rather, it extends it.

6 Reflections

This final chapter of the dissertation framework reflects on the results of the thesis and its underlying research. Chapter 6.1 discusses what can be learned for research at the intersection of computer science and philosophy. Chapter 6.2 discusses limitations and implications for further research.

6.1 Research at the Intersection of Computer Science and Philosophy

As the thesis discusses in-depth, computer technology and computer science evolve constantly. Computer ethics must keep up with these developments and adjust accordingly.[32] The thesis provides examples of how such adjustments can be made. This chapter reflects on what can be learned from these examples (and the thesis more generally) for achieving a productive co-evolution of computer technology, computer science, and computer ethics. It links these observations to literature focusing on research at the intersection of computer science and philosophy, which also serves as a starting point for the following reflections.

On the level of individual case studies, literature on approaches such as VSD already outlines procedures for how methods of computer science, the social sciences, and philosophy can "inform and shape and reshape each other" (Friedman & Hendry, 2019, p. 35).[33] However, it is only recently that more general reflections on methods beyond individual case studies have focused on the potential and need for closer integration of computer science and philosophy.

Friedman et al.'s (2021) argument that aligning approaches such as VSD "with existing innovation and engineering practices" is critical to enabling large-scale industrial adoption is one such example. Another is Winkler and Spiekermann's (2018) call for the provision of more "methodological guidance" and "best practices" by computer ethicists to computer scientists and developers. In line with this perspective, scholars

[32] This is not to be understood as a technology deterministic argument, describing a development of the computer technology as following an inherent logic without any social influence. It merely points out that computer ethics is not the sole driving force behind the development of computer technology and must therefore take into account developments driven by other influences.

[33] See also the description of VSD's iterative approach in Chapter 3.1.4.

working on VSD have started to develop method toolkits that provide more hands-on instructions and practical guidance for VSD practitioners (Friedman et al., 2017; Friedman & Hendry, 2019).

The results of this thesis and the discussions in the IGT and GOAL research projects suggest that such efforts are worthwhile and should be intensified. Especially bringing more expertise from computer science to computer ethics is crucial for adapting these approaches to changing socio-technical realities. The thesis describes two different ways this can be achieved in practice.

First, the thesis shows that the expertise of computer science with respect to specific socio-technical systems should not only be leveraged when *applying* individual approaches to computer ethics. It should also play a more significant role in *refining* such approaches and *selecting* appropriate ones for a specific context. The expertise can refer, for example, to technical features, involved actors, or the interaction between those actors. Especially in the case of more established socio-technical systems,[34] computer ethics can rely on an abundance of existing studies from disciplines such as computer science or information systems. Computer ethics should leverage these sources of knowledge to further develop its approaches and continuously adjust them to emerging or changing requirements.

A study by Umbrello and van de Poel (2021) on the application of VSD in the context of AI systems exemplifies how insights into specific systems can be used to adapt and refine approaches to computer ethics. They argue that "[d]ue to their self-learning capabilities, AI systems (especially those powered by ML) may develop features that were never intended or foreseen – or even foreseeable – by their designers" (Umbrello & van de Poel, 2021, p. 286). Because of this feature of ML-based AI systems, they suggest modifying the VSD approach in the context of these systems to cover their entire lifecycle. Thereby, they aim to prevent constantly evolving AI systems from adapting to new data in ways that "disembody values embedded in their original design" (Umbrello & van de Poel, 2021, p. 286).

Secondly, computer ethics can integrate concepts from computer science and related disciplines such as information systems. Integrating such concepts can help to better

[34] For example, the Bitcoin blockchain or Apple's iOS and App Store.

understand and address challenges to computer ethics. An example of this practice in the thesis is the integration of the boundary resources concept. As Publication 2 shows, insights gained based on this concept can be used to determine how approaches to computer ethics should be applied. For instance, should they be applied to directly promote specific ethical values and principles? Or alternatively, is it more promising to either encourage powerful actors in the system to do so? Or, to (re)shape how power manifests among actors involved in the system, so that (other) approaches to computer ethics can be applied more effectively henceforth (cf. Chapter 5.1.2, Chapter 5.2.2, and Publication 4)?

Further research building on this thesis could, for example, develop or identify concepts analogous to boundary resources for socio-technical systems other than platform ecosystems. Such concepts could help to systematically capture how power manifests in these systems' technical, social, political, and economic features. Especially in the case of emerging (and thus not yet adequately studied) technologies, there may be no suitable existing concepts at hand. In such cases, the collaborative interdisciplinary development of concepts seems promising

Alternatively, existing concepts from the sociology and philosophy of technology can be applied and made tangible for the given context with the help of computer scientists. For instance, the somewhat more abstract concept of boundary objects[35] could be adopted as a more widely applicable alternative to the boundary resources concept (cf. Bowker & Star, 1999; Star & Griesemer, 1989). Together with computer scientists and information systems researchers, the concept could be used to analyze and describe how different actors in a system interact and what constrains their actions. This, in turn, would allow computer ethicists to decide which approaches to computer ethics can be effectively applied, how, and by which actors.

From a computer science perspective, it is thus important to understand approaches to computer ethics not as static or inflexible but rather as dynamic and malleable.

[35] Star and Griesemer (1989, p. 393) define boundary objects as "objects are objects which are both plastic enough to adapt to local needs and the constraints of the several parties employing them, yet robust enough to maintain a common identity across sites. They are weakly structured in common use, and become strongly structured in individual- site use. These objects may be abstract or concrete. They have different meanings in different social worlds but their structure is common enough to more than one world to make them recognizable, a means of translation. The creation and management of boundary objects is a key process in developing and maintaining coherence across intersecting social worlds."

Computer scientists and developers themselves can address some of the challenges practitioners with a more technical background currently have when applying approaches such VSD (see, e.g., Winkler & Spiekermann, 2018). They can do so by providing methodological, conceptual, or technical expertise to refine these approaches and remove barriers to their application. Conversely, computer ethics can proactively invite such input from more technical disciplines and use them to a greater extent in method development and refinement.

6.2 Limitations and Implications for Further Research

As with any research project, the thesis has some limitations. This chapter reflects on these limitations. Furthermore, it outlines how further research could address them.

The first limitation concerns the fact that the thesis has been written cumulatively over a period of several years. Within this timeframe, research perspectives, the examined computer technologies, and the legal environment and related discourses evolved. In particular, the rapidly advancing discourse around EU regulations such as the Digital Markets Act, the Digital Services Act, and the AI Act must be mentioned. These regulatory proposals have been developed and then continuously refined, evaluated, and commented on while the thesis has been written. As a result, the thesis' underlying understanding of the to-be-expected future state of regulation advanced. This is reflected, for example, in different assessments of the potential impact of regulation on computer ethics.

The second limitation concerns the ongoing legislative process of several policy proposals discussed in this thesis. A core theme of this thesis is the analysis of the role of computer ethics in increasingly regulated environments. Some of the assertions made in this thesis rely on the assumption that regulations such as the Digital Markets Act, the Digital Services Act, and the AI Act will have a certain impact. This is because the thesis – among other things – aims to stimulate reflections on the relationship between regulation and computer ethics. While the Digital Markets Act and the Digital Services Act have already entered into force, they will become applicable later in 2023 and 2024, respectively. The AI Act, in contrast, is still in the legislative process. Consequently, there might be changes to the respective regulatory proposals, or they might not have the assumed effects. Yet, the thesis argues that these regulatory proposals are still instructive examples for reexamining some aspects of computer ethics. Nevertheless, in the future, some assessments may prove misguided or relate to aspects that will no longer be part of the respective regulations once these are fully implemented.

Addressing this limitation requires further empirical research to assess the respective regulations' impact. Additional studies could examine whether the Digital Markets Act, once applicable, improves the leverage of app developers against platform providers. Greene and Shilton (2017) could serve as a blueprint for such research. The study analyzes "discussions about privacy on two major developer forums" and how legal guidelines of platforms such as Android or iOS impact these discussions. Similarly, future research could approach how regulations such as the Digital Services Act, Digital Markets Act, and AI Act affect discussions by developers on ethical values more generally. However, such empirical investigations can only begin once these regulations are in place and applicable.

The third limitation concerns the lack of practical evaluation of the thesis' findings concerning design-oriented computer ethics. This limitation arises because this thesis focuses on contexts where applying these approaches is particularly challenging for most actors. However, while writing this thesis, some actors have been identified for whom the application of approaches to design-oriented computer ethics might still be possible or even particularly effective. These include, for instance, platform providers in platform ecosystems (Publication 2) or data brokers who control access to otherwise unavailable training data for ML-based AI systems (Publication 4).

Thus, further research in cooperation with such actors could address this limitation. For example, the potential of approaches such as VSD for designing boundary resources in platform ecosystems could be assessed. Publication 2 suggests that this approach might allow shaping a socio-technical system as a whole, including the design of components not directly controlled by the actor applying VSD. Further studies in cooperation with such powerful actors could empirically evaluate whether such a *value sensitive shaping of ecosystems* (cf. Publication 2) is indeed feasible.

The fourth limitation concerns the thesis' perspective on computer ethics. The thesis focuses on design- and policy-oriented computer ethics specifically. However, there are further approaches to computer ethics. Presumably, some of these approaches are also affected by the developments discussed in this thesis. For example, professional ethics could also be affected by advancing regulations. Here, the question arises of how the norms and rules of professional codes of conduct relate to these regulations. Another area of interest is the extent to which legal and ethical obligations for professionals coincide. Further research could reexamine computer ethics more broadly in light of the developments discussed in this thesis. Reexamining guidelines for professional ethics could be starting point. Such efforts would be in line with current research on AI ethics, reflecting critically on AI ethics guidelines (Munn, 2022; van Maanen, 2022; Wagner, 2019). Establishing additional links between these two strands of literature could contribute to a more holistic view of the emerging challenges and opportunities for AI and computer ethics.

The fifth limitation concerns the thesis' analyses of the actor constellations in the respective socio-technical systems. Here, the focus was in some respects selective and, thus, not exhaustive. In the case of platform ecosystems, for instance, the role of third parties providing software elements or services to developers of applications was not examined in depth. The same applies to users of AI systems, who often also provide data to these systems. Yet, it was necessary to focus on a limited number of actors to be able to take different socio-technical systems into account. Considering different socio-technical systems was crucial because it allowed for a broader reexamination of computer ethics and more general conclusions. Thus, another avenue for further research is to shed light on the role of actors who have received little attention in this thesis.

7 Publication 1:

How Implicit Assumptions on the Nature of Trust Shape the Understanding of the Blockchain Technology

Authored by Mattis Jacobs

Jacobs, M. (2021). How Implicit Assumptions on the Nature of Trust Shape the Understanding of the Blockchain Technology. *Philosophy & Technology*, *34*(3), 573–587. https://doi.org/10.1007/s13347-020-00410-x

Citation style and bibliography harmonized

How Implicit Assumptions on the Nature of Trust Shape the Understanding of the Blockchain Technology

Mattis Jacobs

Abstract

The role that trust plays in blockchain-based systems is understood and portrayed in various manners. The blockchain technology is said to enable and establish trust as well as to redirect it, to substitute for it, and to make it obsolete. Furthermore, there is disagreement on whom or what users have to trust when using the blockchain technology: (only) code, math, algorithms, and machines, or still (also) human actors. This paper hypothesizes that the divergences of the depictions largely rest on implicitly adhering to different accounts of trust. Thus, the goal of this paper is to outline how the current lack of a shared understanding of the term "trust" leads to diverging interpretations of the blockchain technology's core features. Furthermore, it shows how this lack of common understanding obstructs scholars from referring to one another meaningfully in the discourse on blockchain technology. To do so, this paper outlines the most prominent depictions of the setup of relevant trust relationships within blockchain-based systems and traces their roots to different underlying assumptions on the nature of trust.

Introduction

On November 1, 2008, an author or a group of authors under the pseudonym "Satoshi Nakamoto" published a whitepaper titled *Bitcoin: A Peer-to-Peer Electronic Cash System* (Nakamoto, 2008) on a cypherpunk mailing list. It outlined a novel approach for enabling cryptocurrencies, apparently free of centralized authority and without roots in incumbent institutions. The implementation of Bitcoin followed in 2009. In the coming years, it gained significant attention as a proof of concept for "the next step in the evolution of money" (Maurer et al., 2013, p. 273). However, the underlying blockchain[1,2] technology quickly outgrew the application area of cryptocurrencies. Subsequent blockchain-based systems incorporate Turing-complete virtual machines instead of supporting only very limited scripting languages. Thereby, they do not only allow the digital transfer value without relying on the third-party intermediaries (cf. Swan & De Filippi, 2017) but also enable so-called self-executing smart contracts, decentralized applications (DApps), decentralized autonomous organizations (DAOs), and several other novel phenomena and organizational structures (Buterin, 2014).

Thus, as *The Economist* (2015) keenly observed regarding the blockchain technology, "[t]he real innovation is not the digital coins." Instead, the article identifies the unusual role of trust in blockchain-based systems as the outstanding element across all application areas. However, while many scholars share this view (Antonopoulos, 2017; Beck et al., 2016; De Filippi, 2017; Hawlitschek et al., 2018; Mallard et al., 2014; Werbach, 2018), the role that trust plays in these systems is understood and portrayed in various manners. The blockchain technology is said to enable (Underwood, 2016, p. 16) and establish (Krishna, 2015) trust as well as to redirect it (Werbach, 2018, p. 30), to substitute for it (Freeman et al., 2020, p. 69), and to make it obsolete (Nakamoto, 2008, p. 8). Furthermore, there is disagreement on whom or what users have to trust when using the blockchain technology: (only) code, math, algorithms, and machines (Maurer et al., 2013; Nakamoto, 2008), or still (also) human actors (Botsman, 2017; Walch, 2019b; Werbach, 2018). While some depictions of the role of trust in blockchain-based systems prove to be dominant in the discourse, no agreeable and comprehensive one has asserted itself to this day.

This paper hypothesizes that the divergences of the depictions largely rest on implicitly adhering to different accounts of trust. As Hardin (2002, pp. 87–88) notes more generally regarding discourses around trust, if they "are to be understood, [participants in the discourse] must specify more narrowly how [they] mean to use the

[1] If not specified differently, the term "blockchain" in this paper only refers to open, permissionless systems. Furthermore, only direct interactions with blockchain-based systems are taken into account. Because second layer applications, i.e., applications building on top of these systems, do not necessarily share all the relevant features with the systems they are based on, they are not considered in this paper.

[2] In most academic literature, the term "blockchain" increasingly gets supplanted by broader terms like "distributed ledger technology" or "append-only databases." However, these terms include also systems with similar characteristics but different operating principles. Since the trust issues discussed in this paper largely depend on the operating principle, the line of reasoning and the results cannot necessarily be transferred to those systems. Therefore, the term "blockchain" is still applied in this paper.

term.” Thus, the goal of this paper is to outline how the current lack of a shared understanding of the term “trust” leads to diverging interpretations of the blockchain technology’s core features. Furthermore, it shows how this lack of common understanding obstructs scholars from referring to one another meaningfully in the discourse on blockchain technology. To do so, this paper outlines the most prominent depictions of the setup of relevant trust relationships within blockchain-based systems and traces their roots to different underlying assumptions on the nature of trust.

Depictions of the Role of Trust in Blockchain-Based Systems

The Bitcoin Whitepaper does not merely outline the technical foundations of the blockchain technology but also provides an interpretation of the role of trust in it. Nakamoto characterizes the (Bitcoin-)blockchain as trust free, i.e., he suggests users can use it “without relying on trust” (Nakamoto, 2008, p. 8). However, as the discourse matures, the notion of a trust-free technology is often used in a narrower sense and refers only to one of two things. On the one hand, the characterization is used to suggest that the necessity to either trust transactional counterparties or intermediaries vanishes when using the blockchain technology. This necessity usually exists when transferring assets with traditional payment processors. For instance, Swan (2015, p. xii) follows this line of reasoning and describes trust-free transactions as “at its most basic level, intermediary-free transactions.”

On the other hand, there is a temporal dimension. The alleged trust-free nature of the blockchain technology also manifests in its capacity to determine future action. Without the application of blockchain technology, users have to trust external entities to perform certain actions in the future, as, e.g., enforcing contracts or controlling the money supply in a desired way. In contrast, the blockchain technology promises to predetermine such actions. It enables users to create self-enforcing contracts and use a currency that is “produced at a predictable rate, with a maximum number [of tokens] pre-established” (Christopher, 2016, p. 172). Therefore, DuPont and Maurer (2015, p. 9) argue that the blockchain technology “seeks to put boundaries around uncertainty”—a sine qua non of trust—“to the point of snuffing it out.” However, this line of reasoning refers to specific features of the blockchain technology. It does not allege that users of blockchain-based systems do not encounter uncertainties or trust issues at all.

In contrast, the depiction of blockchain as a technology based on trust in code, math, or algorithms suggests that blockchain-based systems do not eliminate the need to trust at all. Instead, it suggests that there is a shift concerning whom—or what—users have to trust when using blockchain-based systems in comparison with whom or what they have to trust when transacting and interacting by other means. Just like advocates of the depiction of blockchain as a trust-free technology, advocates of this depiction also assume that trust becomes obsolete in some areas. For instance, Maurer et al. (2013, p. 264) suggest that for "Bitcoin to work, one does not have to trust Nakamoto, a bank, or any other person or institution." However, in contrast to the depiction of blockchain as a trust-free technology, this depiction of blockchain technology does not end with the determination of where relationships based on trust become obsolete. Instead, it outlines what they are replaced with. Here, Maurer et al. (2013, p. 264) note that instead of trusting intermediaries, one "must simply trust the code or, more precisely, the cryptographic algorithm."

The third depiction assumes that the role of trust in blockchain-based systems differs from trust in other setups in that it is placed in networks of actors instead of individual actors (Werbach, 2017, p. 501). For this co-founder of LinkedIn Reid Hoffman coined the term "trustless trust." He suggests that the setup of trust relationships within blockchain-based systems comprises relationships of a novel nature in which no individual can be identified as the sole trustee (Hoffman, 2014). Another commonly used term to describe this phenomenon is "distributed trust." This insight constitutes the basis for Werbach's seminal book *The Blockchain and the New Architecture of Trust* (2018), potentially the most comprehensive work on the issue so far. Based on the understanding of the blockchain technology as a facilitator of distributed trust, Werbach sheds light on the differences between the role of trust in blockchain-based systems and the role of trust in other predominant setups in society. Here, he references peer-to-peer trust, i.e., interpersonal trust between transacting individuals; "Leviathan" trust, i.e., trust that is established by a "powerful central authority operates largely in the background to prevent others from imposing their will through force or trickery"; and intermediated trust, i.e., trust that is established through the internal rules "and the reputation of the intermediaries" Werbach (2018, p. 27).

Diverging Assumptions on the Nature of Trust

The term "trust" is often insufficiently defined in publications discussing the role of trust in blockchain-based systems. However, trust is a multi-faceted term, and its meaning is not always evident in a given publication. This especially holds true for interdisciplinary contexts. As shown in the following, one reason for the emergence of diverging views on the role of trust in blockchain-based systems is the adherence to different accounts of trust. This section introduces shared as well as conflicting assumptions among various accounts of trust in order to illustrate how using the term in one way or another changes the understanding of the role of trust in blockchain-based systems.

An assumption at the core of most accounts of trust is its basic structure as a three-place predicate of the form A trusts B to/with X*,* where *A* and *B* constitute actors—the trustor (or truster) and the trustee—and *X* and action, testimony, or "valued thing" (Baier, 1986, p. 236; cf. Hardin, 2002, p. 9). Furthermore, trust is a reductive term in the sense that it "is not a primitive, something that we just know by inspection, as the color blue might be a primitive [...]. Rather, it is reducible to other things that go into determining trust" (Hardin, 2002, p. 57). What these "other things" or components of trust are, however, scholars disagree. Most accounts of trust assume that trust requires the trustor to "1) be vulnerable to others [...]; 2) think well of others, at least in certain domains; and 3) be optimistic that they are, or at least will be, competent in certain respects" (McLeod, 2006). *Less demanding* accounts of trust stop here. For instance, Gambetta (1988, p. 217) defines trust in this line of thinking as "a particular level of the subjective probability with which an agent assesses that another agent or group of agents will perform a particular action [...]." Trust, according to less demanding accounts, is a matter of *rational expectations* and is often equated with mere judgments of reliability.[3]

However, another school of thought claims that trust and mere judgments of reliability are disparate concepts. It suggests that trust is a *more demanding* concept. Thus, according to richer, i.e., more demanding accounts of trust, for an attitude to qualify

[3] In fact, trust according to these accounts is most often not compared with judgments of reliability but with reliance. However, as Nickel (2013, p. 224) observes, "this is not a suitable comparison. Reliance is way of acting, whereas trust is an attitude." This paper follows Nickel's reasoning that the appropriate attitudinal counterpart for trust is a "judgment of reliability."

as trust, it needs to meet more requirements. In other words, more demanding accounts of trust define trust as a judgment of reliability + x. However, what the "magic ingredient which distinguishes (dis)trust from mere (non-)reliance" (Hawley, 2017, p. 231) is even scholars who call for the distinction disagree. Candidates are, e.g., certain motivations such as goodwill towards the trustor (Baier, 1986), encapsulated interest, i.e., the idea that trusting requires "a commitment to acting at least partly in the interests of the truster because they are the interests of the truster" (Hardin, 2002, p. 57), the appropriateness of specific reactive attitudes such as feeling betrayed in case of misplaced trust (Baier, 1986), or the necessity of the trustor to see "a trustee as morally obligated, committed, or accountable in appropriate ways" (Hawley, 2017, p. 231; see also Simon, 2013).

The requirements of both more and less demanding accounts of trust are mostly discussed with regard to interpersonal settings, i.e., for trust relationships in which the trustor and the trustee are individual human actors. Accounts of trust that consider other entities, potential trustees are usually defined by outlining their divergence from interpersonal trust as the initial and basic concept. Examples of such accounts are, e.g., accounts of trust in groups and organizations (Hawley, 2017), accounts of trust in governments (Hardin, 2002), or accounts of trust in technological systems (Nickel, 2013). Less demanding accounts of trust based on rational expectations are more readily applicable to non-interpersonal settings.

However, in blockchain-based systems, even concerning the relationships among human actors, it is not clear whether a demanding account of interpersonal trust can be applied. This is due to specific technical features of open and permissionless blockchain-based systems in which users are represented through digital keys. The resulting pseudonymity has the effect that users do not necessarily know with whom they are engaging. This eliminates the possibility to evaluate "contextual features" (Werbach, 2018, p. 29)of contractual counterparts and validators, i.e., miners. This non-identification goes along with an open validation process that does not depend on excluding actors that do not ex ante prove to be trustworthy (Antonopoulos, 2014).

Thus, if the relationship of the two contracting parties or between users and miners is characterized as a relationship based on trust, a less demanding account of trust that does not require a "grounding in specific prior or subsequent relationships with those others" (Hardin, 2002, p. 60) needs to be applied. Hardin (2002, p. 62), who advocates a more demanding account of trust, rejects the idea of such generalized trustees. He alleges that the respective propositions do not really claim "that one trusts those others, but only that one has relatively optimistic expectations of being able to build successful relationships with certain, perhaps numerous, others […]."

Taking one or the other side in the debate on what qualifies as an instance of trust is decisive for how relationships among different human actors in blockchain-based systems are to be characterized. If a more demanding account of interpersonal trust is applied, the relationships between transacting parties as well as between users and miners cannot be considered to be relationships based on trust. Instead, the respective stances could be characterized as mere judgments of reliability. Taking this perspective, the characterization of blockchain-based systems as trust-free appears to be much more tenable, without even touching the issue of whether and, if so, where exactly vulnerabilities and uncertainties vanish. If dealing with the many vulnerabilities and uncertainties existing in blockchain-based systems requires only mere judgments of reliability, then this points to the validity of the argument that trust plays a lesser role in blockchain-based systems compared with other setups. Yet, if blockchain-based systems are indeed entirely trust-free remains to be shown. This issue is addressed again later on.

Similar questions arise regarding whether or not regulatory bodies (i.e., governmental institutions), core developers (i.e., more or less institutionalized groups), the distributed network of miners, and newly established intermediaries like cryptocurrency exchanges (i.e., organizations) qualify as potential trustees. While large proportions of the literature on trust focus on interpersonal trust, some scholars also introduce accounts of trust that are non-interpersonal, "including 'institutional trust' (i.e., trust in institutions), trust in government" (McLeod, 2006) as well as in technological systems (Nickel, 2013), groups and organizations (Hawley, 2017), and many more. Therefore, the accounts of trust in the above-mentioned entities are worth assessing, even though requirements of more demanding accounts stemming from interpersonal settings do not prima facie appear to be applicable.

Less demanding accounts of trust based on rational expectations that equate trust with judgments of reliability are generally open to being applied in contexts where the trustee is not a human actor but, e.g., an organization such as a cryptocurrency exchange. According to such less demanding accounts of trust, it could be sufficient to "be confident that the design of the roles and their related incentives will get role holders to do what they must do if the organization is to fulfill our trust." Still, a reasonable judgment would require having a clear understanding of the structures of and the roles within an institution "to be confident of the incentives or other motivations that foster trustworthiness among role holders" (Hardin, 2002, p. 156). If, on the contrary, a more demanding account is applied, it is questionable whether such entities qualify to constitute a trustee. For instance, Hardin's account, which focuses on encapsulated interest, is incongruous with such entities as trustees. According to his account, the trustor would be required to "know that the agents or the institution act on [the trustors] behalf because they wish to maintain their relationships with" them (Hardin, 2002, p. 156). For larger institutions, he argues that this is "generally not possible."

However, the matter is contested. Hawley (2017) makes the counterargument that the distinction between trust and mere judgments of reliability matter at the individual level but less so at the level of collective entities. Collective entities like institutions—in contrast to individuals—can have an obligation to be reliable. She claims that there "is no general obligation upon individuals to be reliable, which is why we need the language of trustworthiness to highlight those particular respects in which individuals are obliged to be reliable. Nevertheless, we can require of our institutions that they be reliable in the respects that matter to us [...]." Therefore, negative reactive attitudes, such as feeling betrayed, can be reasonably applied in the case that certain institutions prove to be unreliable. This, according to Baier (1986) and others, is only appropriate in contexts where genuine trust is required. Hawley therefore argues for abandoning the distinction between mere judgments of reliability and trust on the level of collective entities. The abandoning of the distinction thus opens the door for a morally laden account of trust regarding organizations and institutions. Accordingly, whether or not groups, organizations, and institutions qualify as trustees varies even among different more demanding accounts of trust.

The answer to the question of whether or not technological artifacts can constitute trustees appears to be more clear-cut. If a more demanding account of trust that exceeds judgments of reliability based on rational expectations is applied, claims of trust in technology can only be understood as trust in the human actors behind the technology (Holton, 1994, p. 66; Jones, 1996, p. 14; Nickel, 2013, p. 224).

Technological artifacts themselves do not have intentions or motivations concerning the actors who assess their reliability. Moreover, actors who assess their reliability usually do not assume this. According to more demanding accounts of trust, technological artifacts are “paradigmatic examples of things about which we make judgments of reliability rather than things we can genuinely trust” (Nickel, 2013, p. 224). Even though less demanding accounts of trust which do not differentiate between trust and mere judgments of reliability mostly stem from the idea of interpersonal trust, they are generally also open for technological artifacts as trustees. If trust is equated with judgments of reliability, technological artifacts are trusted if an actor assesses their reliability positively.

Thus, assumptions on the nature of trust have significant implications on how the role of trust in blockchain-based systems can be conceptualized. Some depictions of “the trust revolution of the blockchain and distributed ledger technology” (Werbach, 2018, p. 30) are evidently only compatible with a specific account of trust, even if the adherence to the account is not made explicit. According to more demanding accounts of trust, fewer stances are considered genuine trust, whereas, according to less demanding accounts of trust, more stances are considered genuine trust. Describing the shift as one from human actors (groups and institutions or individuals) as trustees to “algorithms that govern users’ interactions” (Hawlitschek et al., 2018, p. 57), a “cryptographic algorithm” (Maurer et al., 2013, p. 264), “the instrumental operation of mining” (Velasco, 2017, p. 722), “an open source code” (Atzori, 2015, p. 7), or “collectives of machines” (Werbach, 2018, p. 30) are incongruous with more demanding accounts of trust that presume specific moral or attitudinal components a prerequisite. Trust in the mining community and other collective entities whose members are generally assumed to act exclusively based on self-interest (Werbach, 2017, p. 504) also appear to be incompatible with at least some of the more demanding accounts of trust.

On the Boundaries of Blockchain-Based Systems

The conceptual matters regarding the nature of trust are not a standalone issue. Adhering to different accounts of trust in the assessment of the role of trust in blockchain-based systems requires focusing on different relationships among actors. For instance, basic assumptions of scholarly disciplines can entail specific conceptions of trust. As shown in the previous section, these conceptions determine whether or not specific components of a socio-technical system are considered potential trustees. If components of a specific type, e.g., cryptographic algorithms, are considered potential trustees, a comprehensive analysis of trust relationships among entities in blockchain-based systems requires the incorporation of components of this type in the respective investigation. This section outlines the varying approaches to including and excluding different actors and entities in the assessment of the role of trust in blockchain-based systems.

The most crucial distinction regarding this issue is between, on the one hand, inquiries that look at blockchains as either closed ecosystems or technical models with strictly defined boundaries and, on the other hand, inquiries that consider currently implemented solutions and the broader environment that these solutions are embedded in. The latter inquiries take into account many more actors that users can potentially be vulnerable to, e.g., cryptocurrency exchanges, regulators, developers, and the human actors behind the cryptographic keys that represent users. Furthermore, they also often include interactions that exceed the boundaries of the narrower technical system and thus do not entirely fall in the purview of the technology's security features. Instances of exchanges that exceed the technical boundaries are, e.g., exchanges of on-chain assets (e.g., cryptocurrency tokens) for off-chain assets (e.g., Fiat money). Thus, the stricter the focus is on the narrower technical system, the fewer vulnerabilities and uncertainties are taken into account. Accordingly, the more sensible depictions of the blockchain technology as being trust-free or only based on trust in technological components appear.

More technical literature highlighting the innovative nature of the blockchain technology often has this rather narrow scope and ignores actors such as cryptocurrency exchanges that raise severe trust issues. Examples of such elaborations can be found in the Bitcoin Whitepaper. While Nakamoto's outlining of his motivation to develop Bitcoin sometimes transcends the boundaries of technical modeling and takes societal aspects into account, the statements regarding the Bitcoin blockchain as "a system for electronic transactions without relying on trust" remain relatively strictly within these boundaries (Nakamoto, 2008, p. 8). The same can be said about Antonopoulos (2014) and his depiction of the blockchain technology as being based on trust by computation. His elaborations also stay within the boundaries of technical modeling, and the discussion of the role of trust omits taking the broader environment of blockchain-based systems into account.

However, scholars such as Botsman and Werbach broaden the scope and take the fringes of blockchain-based systems and the (socio-)technical layers underlying them into account. While acknowledging trust minimizing features within the narrower technical system, they corroborate that humans are still very much in the loop in blockchain-based systems and they can exert power individually, i.e., in a non-distributed manner. Botsman (2017) here itemizes "programmers, [...] entrepreneurs and experts who establish and maintain the cryptographic protocols." The list can be complemented by regulators, cryptocurrency exchanges, providers of the underlying internet infrastructure, and many more (see De Filippi & Wright, 2018). Since these actors do not operate according to the blockchain protocol but exert power by other means, potential trust towards them is not based on the game theoretical assumptions underlying the protocol's incentive system.

Thus, if the broader environment of blockchain-based systems is taken into account, issues neglected in a characterization that considers only the actors within the narrower technical system become visible. Regularly, for instance, cryptocurrency exchanges and their users fall as victim to hacks, attacks, and frauds (Chohan, 2018). Since they are for most users next to impossible to circumvent and users are highly vulnerable through and towards them, uncertainties (which one might argue require trust to overcome) to some degree thwart the steps taken within the narrower system to move away from the need to trust intermediaries. Furthermore, if (core-)developers and their abilities to assert changes to the system's protocol are also taken into account, it is necessary to consider the system itself an ever-evolving rather than a static entity. Therefore, especially in the case of transactions with a longer settlement duration, uncertainties resulting from updates of the system's protocol play a prominent role. The maintenance of the features of the technical system themselves, including the ones responsible for the alleged trust-minimizing features of the narrower system, are dependent on human actors, which, as Walch (2019b) argues, users have to trust in turn. Thus, if the broader environment of the technical systems is taken into account, it is accordingly possible to identify trust relationships between users and other (human) actors that can be described in terms of more demanding accounts of trust, i.e., accounts of trust that exceed rational expectations. However, both developers and the broader blockchain community are dedicated to the development of governance mechanisms that limit the capabilities of core-developers to assert power in an uncontrolled manner. Thereby, they reduce the vulnerability of users to developers. Thus, whether or not core developers are (or remain) as powerful as Walch portrays them continues to be a matter of debate.

Furthermore, the perspectives on users vary significantly in the literature on the role of trust in blockchain-based systems. While some inquiries only reflect on the information flow to and from the user and the user's set of possible actions, others also take into account how actual users are able to maneuver within systems. In a common practice in software development, "the 'trusted' label is given to systems that have been tested and proven to have met certain criteria" (Abdul-Rahman & Hailes, 1998, p. 49). These criteria are usually technical. They pay less attention to the capabilities of actual users to maneuver within these systems. However, as Christopher (2016, p. 173) and Greenfield (2017) note, most users do not have the computer literacy necessary to understand and assess the code of their client applications, the blockchain protocol, or particular smart contracts they intend to use. Thus, even if users receive all relevant information necessary to verify certain actions by other users, such as the remittance of funds, they are most likely not able to assess them sufficiently. This phenomenon is not blockchain-specific. It rather emerges in the context of most of the use of modern-day technologies. Nickel (2013, p. 223) therefore points out that "[i]t is impossible for any one person [...] to know enough about how technology works in these different areas to make a calculated choice about whether to rely on the vast majority of the technologies she/he in fact relies upon."

Therefore, most users must consult more computer literate human or institutional actors to assess the various technical subsystems on their behalf as, e.g., computer scientists who assess smart-contract code or cryptocurrency-wallet providers who support the administration of on-chain assets such as cryptocurrency tokens. These actors are not considered in the technical modeling, which largely assumes idealized users who can assess the information given to them. In order to leverage the trust-free or trust-minimizing features of the system that should allow users to not depend on trusting transactional counterparts and third-party intermediaries in a more demanding sense, most users cannot avoid making themselves vulnerable to new actors whom they need to trust in turn.

Thus, there necessarily are limits to the trust-minimizing features of blockchain-based systems. These features have an effect within the boundaries of the narrower technical system but do not fully extend it to actors at the fringes or outside of these boundaries or transactions that transcend them. However, the outlined manifestations of dependencies and vulnerabilities at the fringes of blockchain-based systems are contingent. For instance, the form and importance of these manifestations depend on the social permeation of blockchain-based systems. If cryptocurrency tokens are more widely accepted, the need to cross the boundaries of the system frequently, e.g., to exchange tokens back and forth into Fiat money, could vanish. Moreover, the relevance of cryptocurrency exchanges—one of the most significant factors of uncertainty—could accordingly be diminished substantially.

In this respect, both perspectives have a *raison d'être*. On the one hand, from an engineering perspective, it is reasonable to ignore these contingent factors at the fringes and consider only the technological features that can be impacted by employing means of the discipline. Here, Hawlitschek et al. (2018, p. 59) identify the trust-free properties that also Nakamoto and others describe as manifesting "as long as [the blockchain] operates as a closed ecosystem within its technical boundaries." On the other hand, it is crucial to recognize that—as De Filippi (2018) points out illustratively in the title of one of her articles—"No Blockchain Is an Island." The salient vulnerabilities at the fringes of blockchain-based systems that challenge the depiction of them being trust-free or only based on trust in cryptographic algorithms are worth investigating, especially since they are a significant factor hindering widespread adoption of blockchain-based systems.

The Depictions of the Role of Trust in the Light of these Findings

As shown, the role of trust in blockchain-based systems can be characterized in various ways depending on which account of trust is applied and which components of the socio-technical system are considered essential. This section outlines the relevance of these decisions for the characterization of blockchain-based systems as being trust-free, being based on trust in technological components like cryptographic algorithms, or being based on distributed trust.

To highlight trust-minimizing features of the technology, blockchain advocates often apply a more demanding account of trust that sets higher prerequisites for a stance to be qualified as trust than just being based on rational expectations towards the trustees' behavior. Their positive notion of blockchain as a "trust-free" technology is only comprehensible if "trust-free" refers to the absence of the need to assess specific motivational factors of trustees as, e.g., the benevolence towards the trustor. Since the consensus mechanisms specified in blockchain protocols are based on an incentive system grounded in game theory, they need to be regarded as trust enhancing rather than trust-free if a less demanding account of trust based solely on rational expectation is applied. The notion of trust inherent in alleging that the blockchain technology is trust-free therefore must be read in terms of more demanding accounts of trust. The given incentive systems presuppose that actors are assessed regarding presumed self-interest only (Werbach, 2018, p. 154). Goodwill, benevolence, or encapsulated interests are not taken into account, as required by more demanding accounts of trust. Within this framework, the innovation behind blockchain can be summarized as allowing users to move from having to *trust* institutional actors in a more demanding sense to only having to make *judgments of reliability* of actors based on game-theoretical assumption within a technologically predefined setting.

However, while the depiction of a trust-free technology appears to be generally tenable within this framework, it has been (over-)stretched by blockchain aficionados. "With a zeal bordering on the religious" they "trumpeted the trustlessness"[4] (Christopher, 2016, p. 141) of the systems and declared it "one of the system's core virtues" (Christopher, 2016, p. 172). By neglecting both limitations of the characterization, the application of a more demanding account of trust, as well as a focus limited to the actors within the narrow technical system, they hype an untenable image of a technological system that frees users from most uncertainties and vulnerabilities without recognizing the emergence of new ones. Because some vulnerabilities of the users towards other actors are readily apparent, this narrative gets challenged. Critics here point to actors outside the boundaries of the narrower system. These include regulators (De Filippi & Wright, 2018) and (core-)developers (Walch, 2019b) who can

[4] Christopher (2016) uses the term "trustless" synonymously with how the term "trust-free" is used in this paper.

wield power individually, cryptocurrency exchanges which tend to fall victim to attacks regularly (Chohan, 2018), and the users themselves who often lack the computer literacy necessary to navigate safely within the broader environment of blockchain-based systems (Christopher, 2016; Greenfield, 2017).

However, many of these critiques against the depiction blockchain-based systems as being trust-free apply an account of trust that is less demanding. They do not consider motivational factors like goodwill or encapsulated interest a prerequisite for genuine trust. Thus, ultimately, they do not distinguish between concepts like trust and mere judgments of reliability and adhere to an account that considers only positive predictive expectations. Sometimes, such a rational-expectation-based account of trust is even made explicit (cf. Botsman, 2017). Because it sheds light on existing uncertainties and vulnerabilities of users in the broader environment of these systems, this critical counter-narrative against the exaggerated claims outlined in the previous paragraph is important and well-founded. Yet, it operates with divergent assumptions and is based on a terminology that differs from the ones used in *more reflective* depictions of blockchain as a technology that enables trust-free transactions.

The second depiction of the role of trust in blockchain-based systems suggests that users only have to place trust in algorithms, code, or math. The idea here is that trust is redirected from one trustee to another—from human and institutional actors to allegedly more trustworthy technological artifacts. However, trust in such entities is conceptually very different than interpersonal trust. Thus, suggesting that users are redirecting the same stance—trust—from human actors to technology (or concepts underlying these technologies) neglects that elements of interpersonal trust according to more demanding accounts cannot be modeled on these entities. They lack essential features such as the capacity for goodwill or encapsulated interest. Furthermore, reactive attitudes such as feeling betrayed in case of failed trust cannot be appropriately directed at technological artifacts. Therefore, speaking of trust in these entities is only meaningful if a less demanding account of trust is applied. However, such accounts of trust in technological artifacts or systems exist already (see Nickel, 2013). To make the argument that "a new form of 'algorithmic trust' is created, one that significantly distinguishes itself from the more traditional typology of trust that was initially only between human agents" (Swan & De Filippi, 2017, p. 605), it is necessary to elaborate how this allegedly new form of trust differs from these more generalized notions of trust in technology. So far, this has not been addressed in the respective elaborations.

The third depiction suggests that the blockchain technology makes it possible to replace trust in individual actors with distributed trust. In this line of thinking, the technological components of the system do not constitute the trusted entity. Instead, they enable users to distribute trust over networks of actors without necessarily trusting any individual actor within the network based on given contextual features. Thus, this depiction considers both technical as well as human elements of the socio-technical system. By highlighting the distributed nature of the trustee, it allows distinguishing between the configuration of trust-based relationships in blockchain-based systems on the one hand and a centralized configuration of trust-relationships arranged around an intermediary or a central authority on the other hand.

Other setups containing distributed trust are already familiar from contexts such as accounts of trust in markets, accounts of trust in the wisdom of the crowd, or accounts of trust in reputation systems. These related concepts provide a basic understanding of how the term “distributed trust” can be understood as a meaningful concept, even though the feature of being distributed is incongruous with at least more demanding accounts of interpersonal trust. By pointing at the technological components of the socio-technical system that facilitate the distribution of trust, i.e., the cryptographic algorithms that enable the underlying consensus mechanism, this depiction also gives a clear picture of the innovation behind the blockchain technology. It suggests that the blockchain technology allows, on the one hand, distributing trust where it was hitherto not possible and, on the other hand, distributing trust by means other than those familiar from other contexts.

However, this depiction is only compatible with rather limited accounts of trust, too. The alleged distributed nature of trust does neither allow for the attribution of motivational factors as goodwill or encapsulated interest nor for the plausible application of reactive attitudes such as feeling betrayed in case of failure or breakdown of trust. Botsman (2017) and others make the adherence to a less demanding account of trust explicit by defining trust as, e.g., a “confident relationship with the unknown.” Furthermore, the idea that this setup replaces trust in individual actors is also only tenable within the confines of technical modeling in which actors like (core-)developers are not considered. The role of these actors who have proven to currently have the capability to exert power individually remains unaccounted for in these considerations.

In addition, in their assessment of the role of trust in blockchain-based systems, some scholars (cf. Botsman, 2017; Werbach, 2018) consider not just one but multiple of the depictions introduced in this paper. The strength of these elaborations lies in that they give an overview of the newly established relationships and structural assurances as well as persisting and emerging vulnerabilities within the broader environment of blockchain-based systems. However, this openness often comes at the cost of conciseness. It subsumes very different conceptual stances under the umbrella term "trust." The stances that users have towards a distributed network of miners, towards contractual counterparties, and towards the underlying technical infrastructure vary greatly, even though dealing with uncertainty plays a role in all of them. By applying a very inclusive account of trust that solely focuses on dealing with uncertainty, they lose the conceptual framework to shed light on the differences between the various stances regarding, e.g., what characteristics of the respective trustees trustors are assessing and which moral dimensions some of the relationships might have.

Conclusion

Many scholars agree that one of the exceptional features of the blockchain technology lies in the unusual requirements it sets for users to place trust in other entities within the system when using it. There are undoubtedly various ideas on how the role of trust in blockchain-based systems differs from the role of trust in other setups. As shown, the ambiguity of the term "trust" plays a crucial role here. While most elaborations simply expect the underlying terminology to be self-explanatory, the implicitly underlying accounts vary greatly. Nevertheless, even though some depictions of the role of trust in blockchain-based systems appear to be incongruous with others, most of them cannot be rejected offhand. In a benevolent interpretation, there are accounts of trust that allow all the entities suggested as trustees at the core of the setup to be covered meaningfully. However, some of the depictions contribute more to the understanding of how blockchain-based systems work and what the critical issues are. Especially the depiction of blockchain as a technology that enables technologically facilitated trust in a distributed network of actors that are not trusted individually can be positively highlighted here. Contrary to other depictions, it allows considering both technical components as well as key actors simultaneously.

The heterogeneous terminologies in the respective lines of argumentation make it increasingly challenging for scholars referencing one another in the academic discourse. Propositions that are made—at least implicitly—based on one account of trust are often attacked from a perspective based on a different conceptual and terminological basis. The most striking example here is the debate on whether or not the blockchain technology enables trust-free transactions. As shown, a reasonable argument can be made for this within the limited purview of technical modeling and based on a more demanding account of trust. However, critics of this notion as well as blockchain aficionados often ignore the limitations of this argument and treat the idea of trust-free transactions as an alleged system feature that without more ado can be leveraged in the everyday usage of implemented systems. Here, the lack of a shared terminology contributes to both an unwarranted critique of the initial argument and an unwarranted hype surrounding alleged features the blockchain technology.

In spite of this, the various affected scholarly disciplines and traditions of thought do not allow a uniform underlying account of trust to be defined. To tackle these issues nevertheless, Hardin (2002, p. 87) suggestion that discourse participants need to specify their terminology regarding trust should be taken more seriously. Thus, since a shared terminology across the multitude of involved disciplines does not appear to be an attainable goal, it is paramount that scholars reflect on divergent views and make underlying assumptions and concepts explicit. In the scholarly context, this should be considered both in the formulation and presentation of arguments as well as in academic evaluation processes.

For this purpose, scholars can fall back on comprehensive works that provide overviews over various accounts of trust in a more general sense (Hardin, 2002; McLeod, 2006; Simon, 2013) and on accounts of trust in specific entities and domains. In the context of the discourse on blockchain, e.g., interpersonal trust (Baier, 1986; Hardin, 2002), trust in groups and organizations (Hawley, 2017), trust in technological systems (Nickel, 2013), and trust in game-theoretical settings (Gambetta, 1988; Voss & Tutic, 2020) are particularly noteworthy. Here, future research could build on the findings of this paper by providing scholars with a taxonomy of the accounts of trust relevant in blockchain research.

If these overviews do not provide accounts that suit a specific proposition or argument, e.g., in cases where scholars allege that blockchain establishes a *new* form of trust (cf. Swan & De Filippi, 2017; Werbach, 2018), it is nevertheless necessary to introduce an explicit definition of trust. Here, the fact that trust is a reductive term can be utilized as a starting point for the explication of hitherto implicit assumptions on the nature of trust. Because trust "is reducible to other things that go into determining trust" (Hardin, 2002, p. 57), these "other things" or components can be pointed out individually. For instance, in the case of the depiction of blockchain as a technology that allows for trust-free transactions, this approach requires clarifying that it is the consideration of specific motivations—which allegedly are prerequisites of trust—that become negligible within the boundaries of the narrower technical system.

Depending on the degree of divergence from existing accounts of trust, these new conceptions are also a worthy subject for philosophical investigations and further research. Particularly the notion of distributed trust in the relationship of users and miners enabled through specific technical features which establish an economic incentive system appears to be noteworthy in this regard. Based on the works of scholars such as Antonopoulos (2014, 2017), Hawlitschek et al. (2018), and Werbach (2018), who describe the relationship of trust and technical features in more detail, comparisons to adjacent phenomena such as trust in markets can be drawn to develop an analogous concept. These novel accounts would complement the aforementioned taxonomies.

Acknowledgments

I acknowledge financial support from Hamburg Ministry of Science, Research and Equality in the project Information Governance Technologies under the reference LFF-FV 34. Open access funding provided by Projekt DEAL.

Publication's References

Abdul-Rahman, A., & Hailes, S. (1998). A distributed trust model. In Proceedings of the 1997 Workshop on new security paradigms. Symposium conducted at the meeting of ACM.

Antonopoulos, A. (2014). Bitcoin security model: Trust by computation. Forbes. com, February, 20. Retrieved from http://radar.oreilly.com/2014/02/bitcoin-security-model-trust-by-computation.html

Antonopoulos, A. (2017). Mastering Bitcoin: Programming the open Blockchain (Second ed.). Sebastopol, CA: O'Reilly Media.

Atzori, M. (2015). Blockchain technology and decentralized governance: Is the state still necessary? SSRN Electronic Journal. Advance online publication. https://doi.org/10.2139/ssrn.2709713

Baier, A. (1986). Trust and antitrust. Ethics, 96(2), 231–260. https://doi.org/10.1086/292745

Beck, R., Czepluch, J. S., Lollike, N., & Malone, S. (2016). Blockchain-the gateway to trust-free cryptographic transactions. In ECIS.

Botsman, R. (2017). Who can you trust? How technology brought us together and why it might drive us apart (first edition (eBook)). New York: PublicAffairs.

Buterin, V. (2014). DAOs, DACs, DAs and More: An Incomplete Terminology Guide. Retrieved from https://blog.ethereum.org/2014/05/06/daos-dacs-das-and-more-an-incomplete-terminology-guide/

Chohan, U. (2018). The problems of cryptocurrency thefts and exchange shutdowns. SSRN Electronic Journal. Advance online publication. https://doi.org/10.2139/ssrn.3131702

Christopher, C. M. (2016). The bridging model: Exploring the roles of trust and enforcement in banking, Bitcoin, and the blockchain. Nevada Law Journal, 17, 139.

De Filippi, P. (2017). In Blockchain we Trust: Vertrauenslose Technologie für eine vertrauenslose Gesellschaft. In Rudolf-Augstein-Stiftung (Ed.), *edition suhrkamp: Vol. 2714. Reclaim Autonomy: Selbstermächtigung in der digitalen Weltordnung* (pp. 53–81). Suhrkamp.

De Filippi, P. (2018). *No Blockchain Is an Island*. https://www.coindesk.com/no-blockchain-island/

De Filippi, P., & Wright, A. (2018). *Blockchain and the law: The rule of code*. Harvard University Press.

DuPont, Q., & Maurer, B. (2015). Ledgers and law in the Blockchain. Kings Review (23 June 2015) http://kingsreview.co.uk/magazine/blog/2015/06/23/ledgers-and-law-in-the-blockchain

Freeman, S., Beveridge, I., & Angelis, J. (2020). Drivers of digital trust in the crypto industry. In M. Ragnedda & G. Destefanis (Eds.), Routledge studies in science, technology and society. Blockchain and Web 3.0: Social, economic, and technological challenges (pp. 62–77). London: Routledge.

Gambetta, D. (1988). Can we trust trust? In D. Gambetta (Ed.), Trust: Making and breaking cooperative relations (pp. 213–237). Oxford: Blackwell.

Greenfield, A. (2017). Radical technologies: The design of everyday life. London, New York: Verso.

Hardin, R. (2002). Trust and trustworthiness. The Russell Sage Foundation series on trust: Volume 4. New York: Russell Sage Foundation. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&AN=1069635

Hawley, K. (2017). Trustworthy groups and organizations. In P. Faulkner & T. Simpson (Eds.), The Philosophy of Trust (pp. 230–250). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198732549.003.0014

Hawlitschek, F., Notheisen, B., & Teubner, T. (2018). The limits of trust-free systems: A literature review on blockchain technology and trust in the sharing economy. Electronic Commerce Research and Applications, 29, 50–63. https://doi.org/10.1016/j.elerap.2018.03.005

Hoffman, R. (2014). The Future of the Bitcoin Ecosystem and "Trustless Trust": Why I Invested in Blockstream. Retrieved from https://www.linkedin.com/pulse/20141117154558-1213-the-future-of-the-bitcoin-ecosystem-and-trustless-trust-why-i-invested-in-blockstream

Holton, R. (1994). Deciding to trust, coming to believe. Australasian Journal of Philosophy, 72(1), 63–76. https://doi.org/10.1080/00048409412345881

Jones, K. (1996). Trust as an affective attitude. Ethics, 107(1), 4–25. https://doi.org/10.1086/233694

Krishna, A. (2015). Blockchain: It Really is a Big Deal. Retrieved from https://www.ibm.com/blogs/think/2015/09/blockchain-really-big-deal/

Mallard, A., Méadel, C., & Musiani, F. (2014). The paradoxes of distributed trust: Peer-to-peer architecture and user confidence in Bitcoin. Journal of Peer Production. (4), 1–10.

Maurer, B., Nelms, T. C., & Swartz, L. (2013). "When perhaps the real problem is money itself!": The practical materiality of Bitcoin. Social Semiotics, 23(2), 261–277. https://doi.org/10.1080/10350330.2013.777594

McLeod, C. (2006). Trust. Retrieved from https://plato.stanford.edu/entries/trust

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Retrieved from https://bitcoin.org/bitcoin.pdf

Nickel, P. J. (2013). Trust in Technological Systems. In M. J. Vries, S. O. Hansson, & A. W. M. Meijers (Eds.), Philosophy of engineering and technology, Norms in technology (Vol. 9, pp. 223–237). Dordrecht: Springer.

Simon, J. (2013). Trust. In D. Pritchard (Ed.), Oxford bibliographies. New York: Oxford University Press. https://doi.org/10.1093/obo/9780195396577-0157

Swan, M. (2015). Blockchain: Blueprint for a new economy. Beijing: O'Reilly.

Swan, M., & de Filippi, P. (2017). Toward a philosophy of Blockchain: A symposium: Introduction. Metaphilosophy, 48(5), 603–619. https://doi.org/10.1111/meta.12270

The Economist (2015). The trust machine. Retrieved from https://www.economist.com/leaders/2015/10/31/the-trust-machine

Underwood, S. (2016). Blockchain beyond bitcoin. Communications of the ACM, 59(11), 15–17. https://doi.org/10.1145/2994581

Velasco, P. R. (2017). Computing ledgers and the political ontology of the Blockchain. Metaphilosophy, 48(5), 712–726. https://doi.org/10.1111/meta.12274

Voss, T., & Tutic, A. (2020). Trust and game theory. In J. Simon (Ed.), The Routledge handbook of trust an philosophy. Routledge.

Walch, A. (2019). In code(rs) we trust: Software developers as fiduciaries in public Blockchains. In I. Lianos, P. Hacker, S. Eich, & G. Dimitropoulos (Eds.), Regulating Blockchain: Techno-Social and Legal Challanges (pp. 58–81). Oxford University Press.

Werbach, K. (2017). Trust, but Verify: Why the Blockchain needs the law. Berkeley Technology Law Journal.

Werbach, K. (2018). The blockchain and the new architecture of trust. Information policy series. Cambridge, Massachusetts, London, England: The MIT Press.

8 Publication 2:

Value Sensitive Design and power in socio-technical ecosystems

Authored by Mattis Jacobs, Christian Kurtz, Judith Simon, and Tilo Böhmann

Jacobs, M., Kurtz, C., Simon, J., & Böhmann, T. (2021). Value Sensitive Design and power in socio-technical ecosystems. *Internet Policy Review, 10*(3). https://doi.org/10.14763/2021.3.1580

Citation style and bibliography harmonized

Value Sensitive Design and power in socio-technical ecosystems

Mattis Jacobs, Christian Kurtz, Judith Simon, Tilo Böhmann

Abstract

Recent European policy papers call for the consideration of human values in the design of information technology. Value Sensitive Design (VSD) provides a framework for systematically accounting for values in the design of technical artefacts. This paper examines how the distribution of power within socio-technical ecosystems poses a challenge for the application of VSD. It identifies four crucial factors determining the effect of the distribution of power on VSD: the level of decentralisation of the ecosystem; if VSD is applied at the core or periphery; when power can be exercised (temporality); and the phase of VSD (conceptual, empirical, and technical) that power can be exercised in. Based on these factors, it outlines how the challenge of accounting for power can be addressed.

1. Introduction

Recent European policy papers call for the consideration of human values[1] in the design of information technology (Datenethikkommission, 2019; European Commission, 2019, 2020a; HLEG-AI, 2019). Approaches such as Value Sensitive Design (VSD) promote the idea that human values can be accounted for in the development of technological artefacts and provide a framework for systematically analysing, weighing, and operationalising them (Friedman et al., 2008; Friedman & Hendry, 2019). However, while VSD is well received in the academic context and attracts attention from various disciplines such as computer science and information systems (Friedman et al., 2008; Friedman & Hendry, 2019; Mueller & Heger, 2018;

[1] This paper uses Friedman and Hendry (2019, p. 24) working definition of the term "human value," referring to "what is important to people in their lives, with a focus on ethics and morality". While the term has been criticised as being both under- and over-defined, it provides an appropriate balance for the practical application in the context of Value Sensitive Design. For an in-depth discussion of existing critique on the definition as well as the advantages and disadvantages of various alternative definitions see Friedman and Hendry (2019) and Brey (2010).

Winkler & Spiekermann, 2018), computer ethics (Brey, 2010; Introna, 2005), healthcare (Walton & DeRenzi, 2009), urban design (Borning et al., 2008; Waddell et al., 2008) and others, there are challenges barring the path to widespread adoption.

Reflecting on such "grand challenges", Friedman and Hendry (2019, p. 176) name "accounting for power" as one of them.[2] In this paper, the term "power" refers to "the ability of agents [...] to realize a certain outcome" (Brey, 2008, p. 75)[3]—specifically design decisions—"even against resistance" (Weber, 2019, p. 134). Friedman and Hendry (2019, p. 176) elaborate on the challenge: VSD "has not explicitly addressed how to handle differences in power among [...] stakeholders [and how] best to account for power relations within a value sensitive design framing remains an open question."

This paper argues that this challenge is exacerbated in many cases by the integration of technical artefacts in increasingly vast and complex socio-technical ecosystems defined as "a dynamic community of competing and interdependent people, organizations, and computing systems operating in a complex, capricious environment" (McConahy et al., 2012, p. 1). In such socio-technical ecosystems, the power over the variety of independent design decisions which, in their totality, define the shape of an artefact is often distributed over various actors. Furthermore, due to their socio-technical nature, containing a "technical, social, political, and economic" (van House, 2004, p. 18) as well as "organizational [...] and business" (Feiler et al., 2006, p. 27) domain, power can manifest in various forms in these ecosystems.[4] In order to account for all the domains in which power can manifest itself, the paper adopts a systemic view on power, which "regards power as the property of broader social, economic, cultural, and political networks, institutions, and structures" (Sattarov, 2019, p. 20) and focuses on how "systems confer differentials of

[2] The identification of the "grand challenges" for Value-Sensitive Design that Friedman and Hendry refer to took place at two workshops in 2015 and 2016 organised by Batya Friedman, David Hendry, Jeroen van den Hoven, Alina Huldtgren, Catholijn Jonker, Aimee van Wynsberghe, and Maike Haarbers in Aarhus, Denmark, and Leiden, The Netherlands, respectively. The workshops aimed at "Charting the next decade" for the approach (Friedman & Hendry, 2019).

[3] Accordingly, the conceptualisation of power used in this paper does not capture the ability to exercise control over other agents (Brey, 2008).

[4] See, e.g., van Dijck et al. (2018) for political and economic manifestations of power, Shilton and Greene (2019) for technical manifestations of power, and De Filippi et al. (2020) for social manifestations of power.

dispositional power on agents, thus structuring possibilities for action" (Haugaard, 2010, p. 425; see also Sattarov, 2019).

The remainder of this paper explores the following questions: 1) how and to what extent does the "grand challenge" of accounting for power in VSD get exacerbated by the integration of technical artefacts in increasingly vast and complex socio-technical ecosystems; 2) how does the organisational structure of socio-technical ecosystems affect the challenge of accounting for power in VSD; 3) how can this challenge be addressed; and 4) are there positive effects for VSD if the approach is applied in settings in which the power to make design decisions is distributed over various actors.

Section 2 provides an overview of VSD. Section 3 further elaborates on the actors involved in developing technical artefacts in different types of socio-technical ecosystems and their respective leverage over how developers can account for specific values. It discusses two exemplary types of socio-technical ecosystems that differ in the degree of decentralisation and, thus, the way power is distributed within them: platform ecosystems such as Apple's iOS ecosystem (section 3.1) and blockchain-based systems such as Bitcoin and Ethereum (section 3.2). Section 4 outlines how adopting a power-sensitive ecosystem perspective can foster a more pronounced understanding of the challenge of accounting for power. Based on these observations, section 5 derives some reference points for addressing the challenge of accounting for power in socio-technical ecosystems. Lastly, section 6 concludes.

2. The Value Sensitive Design approach

According to Friedman and Hendry (2019, p. 3), VSD "seeks to guide the shape of being with technology". Directed at "researchers, designers, engineers, policy makers, and anyone working at the intersection of technology and society [...], it provides theory, method, and practice to account for human values in a principled systematic manner throughout the technical design process". To account for values in the design process of technical artefacts, VSD is structured in three phases of action: conceptual, empirical, and technical investigations. Pertinent literature maps out the respective phases primarily by determining what practitioners should aim to achieve in them. The approach deliberately refrains from prescribing specific methods in the individual phases, allowing VSD practitioners to select and integrate methods tailored to the respective context of application on a case-by-case basis (Friedman et al., 2008; Friedman & Hendry, 2019).

Conceptual investigations "comprise analytic, theoretical, or philosophically informed explorations of the central issues and constructs under investigation" (Friedman & Hendry, 2019, p. 12). They address issues such as the identification of direct and indirect stakeholders, the nature of the respective stakeholder's implication, the conceptualisation of values, and dealing with value conflicts (Friedman et al., 2008). Regarding value conflicts, it is important to note that such conflicts can also exist between human and instrumental values (Friedman & Hendry, 2019) and between values that are directly affected and values whose preservation is being put at risk only in the future (see Czeskis et al., 2010). In a more recent publication, Friedman and Hendry (2019) also include developing a framework for evaluating a successful application of VSD into this phase.

Empirical investigations employ quantitative and qualitative social sciences methods to determine the stance of (groups of) stakeholders towards values and their respective weighing (Simon, 2016). Additionally, practitioners can deploy empirical methods in a later stage to "evaluate the success of a particular design" with regards to whether it supports the realisation of a particular value as intended (Friedman et al., 2008, p. 72). Empirical investigations of this second form aim to answer whether the objectives defined in the conceptual investigation have been achieved. In both early- and late-stage empirical investigations, developers apply advanced survey and interview methods to disclose discrepancies between the espoused practice of stakeholders with their actual practice (Friedman et al., 2008). Thus, empirical investigations provide "a more situated understanding of the socio-technical system" in question and facilitate "the observation of stakeholders' usage and appropriation patterns, but also whether the values envisioned in the design process are fulfilled, amended, or subverted" (Simon et al., 2020, p. 4).

Technical investigations also take two different forms in VSD. The first form comprises a retrospective analysis of existing technological artefacts and aims at disclosing "underlying mechanisms [that] support or hinder human values" (Friedman et al., 2008, p. 73). This form thus corresponds roughly to what Brey (2010) and Introna (2005) refer to as "Disclosive Computer Ethics". The second form of technical investigations "involve[s] the proactive design of systems to support values identified in the conceptual investigation" (Friedman et al., 2008, p. 73). Practitioners here address how the respective conceptualisation and weighing of values can be operationalised and accounted for in the design process, i.e., how they can be translated into code.

The three phases of Value Sensitive Design repeat iteratively. Neither a starting point nor an order is prescribed. The respective phases are intended "to inform and shape and reshape each other" through the iterations (Friedman & Hendry, 2019, p. 35).

Because VSD does not prescribe the use of specific methods in the respective phases, recent overview articles and literature reviews (Friedman et al., 2017; Friedman & Hendry, 2019; Winkler & Spiekermann, 2018) provide practitioners of VSD with heuristics on how to proceed by invoking exemplary case studies. They instance methods such as stakeholder analyses (Friedman, Kahn, et al., 2006), value scenarios (Nathan et al., 2007), ethnographically informed inquiries (Nathan, 2012), multi-lifespan timelines (Yoo et al., 2016), and others. Additionally, pertinent literature provides further heuristics such as lists "of human values with ethical import that are often implicated in system design" as a tangible basis for practical application (see also Friedman et al., 2008; Simon et al., 2020, p. 4).

3. Accounting for power in socio-technical ecosystems

In contrast to the development of independently working, stand-alone, monolithic technical artefacts, the development of artefacts integrated into socio-technical ecosystems have a much more constrained scope for design. To enable coordination among providers of components of a socio-technical ecosystem, the component's design must account for the existing technical and non-technical features of the ecosystem. Thus, the actors (co-)determining these features set restrictions to design decisions for novel components of an ecosystem (McConahy et al., 2012), including attempts to account for human values.

The following sections showcase two types of socio-technical ecosystems, characterised by different actor constellations, to reveal the specific manifestation of the challenges for applying VSD in the respective contexts. Section 3.1 focuses on platform-based ecosystems such as Apple's iOS, Google's Android, and the Facebook ecosystem, whereas section 3.2 focuses on blockchain-based ecosystems. The cases differ in how power is distributed in the respective types of ecosystems, how it manifests, how visible its distribution is, and the available modes of governance to address power-related issues. The dissimilarity of the cases allows to take a wide-angle perspective and to identify various facets of the challenge of accounting for power in VSD.

3.1 Platform-based ecosystems

The majority of digital interactions occur in ecosystems, often facilitated and connected over a digital platform (Lusch & Nambisan, 2015). The term "platform" is often a source of confusion due to a variety of definitions. In this paper, the term refers to a software-based system as the core with an extensible codebase that enables functionality for users through additional software subsystems in the form of peripheral applications—or modules—that interoperate with it (Baldwin & Woodard, 2009; Reuver et al., 2018; Tiwana et al., 2010). Peripheral services are provided by developers to enable the provision of functionality, service, or content (Constantinides et al., 2018), which can be accessed by the user via the platform (Ghazawneh & Henfridsson, 2013).

Previous research has already addressed some challenges for applying VSD (or accounting for human values more generally) in platform-based ecosystems (Shilton & Greene, 2016, 2019; van Dijck et al., 2018; Warnier et al., 2015). For instance, focusing on deliberations of mobile application developers in developer forums on how to account for values, especially privacy, Shilton and Greene (2016) demonstrate the extent to which platform providers can assert their ideas without actively engaging in design processes. Shilton conceptualises the platform provider's means to do so as "value levers" (Shilton, 2012) and, together with Greene, provides comparative studies on the use of these levers in different platform ecosystems (Shilton & Greene, 2019). However, a more elaborate ecosystem perspective—as developed in information systems research—could provide a means to refine such an analysis.

According to information systems literature, the platform provider's central role in platform-based ecosystems is a facilitating one (van Alstyne et al., 2016). To scale a platform, the platform provider needs to attract external actors into the platform ecosystem that engage in interactions. In platform-based ecosystems, mainly actors of four different groups come together: users, platform providers, app providers, and third parties. Platforms enable external developers to contribute to the ecosystem by providing so-called *boundary resources* (Constantinides et al., 2018; Tiwana & Konsynski, 2010). Boundary resources are socio-technical manifestations of the platform provider's power to influence a platform ecosystem (Ghazawneh & Henfridsson, 2013), such as application programming interfaces (API), software development kits (SDK), legal guidelines, and application approval processes (Eaton et al., 2015; Karhu et al., 2018).

As control points for a platform provider, boundary resources facilitate an arm's length relationship between the platform provider and service providers (Ghazawneh & Henfridsson, 2013). They offer the providers of peripheral applications access to a platform's resources while allowing the platform provider to retain influence over the platform (Eaton et al., 2015). Using boundary resources, a platform provider orchestrates its platform ecosystems and enables service providers to participate in and contribute to the platform's development (Eaton et al., 2015). Designing and implementing boundary resources is a balancing act of retaining power while supporting service providers to create independent platform-based innovation (Eaton et al., 2015). Thus, platform providers hold the privileged position to exercise power by determining the design of boundary resources and thereby influence the actions of service providers and third parties involved with the platform (Eaton et al., 2015; Ghazawneh & Henfridsson, 2013), with direct implications for how these actors can account for values.

This indicates that platform providers can serve a decisive role in encouraging (or discouraging) design decisions that support the realisation of human values. For instance, in the redesign of its boundary resources via iOS 13, Apple introduced fine-grained user configuration options regarding the usage of location data by apps (Apple, 2019). In previous iOS versions, users could choose among the three options 'Never', 'While Using the App', and 'Always' (Apple, 2019). iOS 13 introduced the additional option 'Ask Next Time'. Users and the developers of applications in Apple's ecosystem are directly affected by such decisions. The configuration options have a considerable impact on user information privacy since the user can make case-by-case decisions on whether or not to grant access to their location data to an app. App developers, on the other hand, have to consider these case-by-case decisions in the expected user behaviour. Platform providers mostly prescribe such changes in boundary resource design unilaterally. Users and developers of peripheral applications are often regarded as passive recipients of these changes. Although the developers of peripheral applications can generate some degree of pressure through public criticism (Hestres, 2013) or building coalitions to achieve their goals (Perez, 2020), the decision-making power lies with the platform provider, in this case, Apple.

In another instance, Apple changed the data interface design for apps to access the MAC (Media Access Control) addresses of the devices an iPhone is connected to. Various applications misused this interface to bypass restrictions on location data access. They approximated the location data by using these MAC addresses in combination with publicly available databases that offered the specific locations of the devices that hold the respective MAC addresses. With the update to iOS 11 in 2017, access to these network data was disabled (Butts, 2017). However, the interface design also blocked data access for app providers that offer network services. As a consequence, these apps were no longer functioning. Thus, Apple restricted the scope for design of app providers, ruling out an operationalisation of privacy that would maintain the existing functionality.

Lastly, digital service providers and related services may also be influenced by the necessity of maintaining and keeping up with platform updates by the platform provider, such as APIs or framework refinements (Ausloos & Veale, 2020). For instance, in OS 14, Apple established the framework *App Tracking Transparency,* playing out privacy features more prominently than in earlier versions of iOS. Due to these changes, app developers have to request user authorisation to access app-related data for tracking the user or the device (Apple, 2020). In the future, Apple intends to ban applications that track users without permission and thus violate the new requirements and respective guidelines (Leswing, 2021). In consequence, Apple's decision to establish third-party transparency has a significant influence on how developers of peripheral applications can conceptualise and operationalise data protection and information privacy.

3.2 Blockchain-based ecosystems

A blockchain is a distributed, encrypted, chronological database of transactions recorded by a distributed network of computers (Morabito, 2017; Wright & De Filippi, 2015). It contains "every transaction that has been carried out and shared among those participating in the network" (Morabito, 2017, p. 4). The entries are "encrypted and organized" in "smaller datasets referred to as 'blocks,'" each of which references "to the preceding block in the blockchain" (Wright & De Filippi, 2015, p. 7). A consensus mechanism warrants the integrity of each transaction over the network. Contrary to other approaches in computer security, in open, permissionless blockchains, the consensus mechanisms are not based on access control, i.e., on "carefully vetting participants and excluding bad actors" (Antonopoulos, 2014). Instead, they rely on economic incentive systems that aim at motivating actors—referred to as miners (Alsindi & Lotti, 2021)—to participate in the validation process and ensuring that it is "more profitable and attractive [for them] to contribute to the network than to attack it" (Brekke & Alsindi, 2021, p. 2). As a result of this approach, "the key characteristics of a blockchain […] are that it is: distributed, decentralized, public or transparent, time-stamped, persistent, and verifiable." (DuPont & Maurer, 2015, p. 2). Moreover, the blockchain technology is not restricted to the record-keeping function it utilises in its origin in cryptocurrencies. More recently developed blockchain-based systems such as Ethereum incorporate Turing-complete virtual machines that allow executing not only simple transactions but also more complex operating steps. In turn, this enables running decentralised second-layer applications (DApps) as services on top of the system.

In line with discourses around earlier decentralised technical systems such as the internet (Bodó et al., 2021), developers, scholars, and the broader community discussed blockchain technology in value-related terms from the very beginning. In this discourse, a libertarian reading of the technology is dominant (De Filippi, 2017; Werbach, 2018; Wright & De Filippi, 2015). Furthermore, trustworthiness (Becker & Bodó, 2021; Hawlitschek et al., 2018; Jacobs, 2021; Werbach, 2018) and sustainability (Alsindi & Lotti, 2021; Giungato et al., 2017) are discussed prominently as values that should be accounted for in the technology's technical design. Notably, design decisions regarding comparably subtle changes in a system's protocol—such as a change in the number of transactions aggregated in one block—are debated by the community in terms of values embodied in the respective design decision (Werbach, 2018).

The technical properties of blockchain-based systems affect the applicability of VSD. In contrast to the platform-based ecosystems discussed in the previous section, there is no central entity controlling the system that can unilaterally determine the design of technical interfaces (Antonopoulos, 2017). Thus, consent between several (groups of) actors is necessary to implement protocol amendments successfully. These are, first and foremost, the developers themselves, but also a significant share of miners, cryptocurrency exchanges, and token holders.

The software protocols of open and permissionless blockchains are maintained as open source projects, i.e., the code is publicly available, everyone can propose or recommend code changes and amendments, and a "mix of volunteer and paid software developers write and update the software" (Walch, 2019b, pp. 60–61). However, while there are no explicit boundaries to participating in the design process in many blockchain-based systems, there are still groups of core developers with additional rights that guide and oversee the design processes in most larger systems (Walch, 2019b). Thus, as Werbach (2018, p. 108) notes, "developers have more power than they let on. [...] And even in an open-source project, a single individual can exercise significant authority."

However, core developers only propose updates. Ultimately, the actors running the network need to "adopt and run [these] implementation[s]" (Antonopoulos, 2017, p. 259). Since the developers do not operate the system, they only create a new version of the system's protocol in a software repository, i.e., they create a "software fork". The respective nodes, miners, and wallet-holders individually decide whether or not they use client software with the updated version of the protocol, i.e., create a "network fork" (Antonopoulos, 2017).

As many protocol upgrades lead to consensus rules that are not "forward compatible" (Antonopoulos, 2017; see also Swan, 2015), i.e., they are incompatible with the pre-upgrade version's ones, miners continuing to proceed according to the old rules and miners proceeding according to the new rules from this point on participate in diverging ledgers. In such a *hard fork*, "two chains evolve independently" from one another (Antonopoulos, 2017, p. 257). If, in the long run, either all or no miners follow the core-developers advice to adhere to the new version of the protocol, one respective branch of the fork perishes. If a sufficiently large group of miners adheres to either version of the protocol, the network splits, with both ledgers persisting. These share the same history but are henceforth dissociated from one another (Antonopoulos, 2017; for a discussion of several cases see DuPont, 2019).

Furthermore, cryptocurrency exchanges, too, need to adopt the new rules for them to be successfully introduced (Antonopoulos, 2017). While exchanges do not directly engage in maintaining or running blockchain-based systems and exist merely at their fringes, they nevertheless impact the incentives that drive the more central actors. For instance, cryptocurrency exchanges need to decide which ledger they list, i.e., for which of the ledgers they offer exchanges to customers and thus provide easy access to the system as a whole. While a delisting on one exchange platform potentially only has negligible effects, a coordinated effort to delist a system by several major platforms can hinder access to the system and diminish the economic incentives to participate in it (Orcutt, 2019).

The actors involved in blockchain-based systems thus do not only potentially hold different values or have diverging preferences and incentives regarding the conceptualisation or operationalisation of values but wield sufficient power to unilaterally intervene in design decisions that concern the realisation of a specific value. Moreover, various recent examples showcase that this is not a mere theoretical possibility, but that diverging preferences regarding values in practice do entice these actors to make use of these means.

Regarding core developers, one of the most prominent instances took place in the aftermath of a hack—commonly referred to as “TheDAO hack”—in which an attacker was able to gain hold of assets worth around “$55 million at the time” (De Filippi & Wright, 2018, p. 141) and siphon them to a fund under its control. However, before the attacker was able to move the assets further or sell them on a cryptocurrency exchange (Botsman, 2017), the core developers of Ethereum pushed through a code update to ultimately void the illicit transactions and “recover the funds from the attackers” (De Filippi & Wright, 2018, p. 141). Because this update affected not just the general principles and functionality of the technology but also individual transactions, commentators commonly use the example of this update to illustrate the power of core developers in the system (De Filippi & Wright, 2018; Walch, 2019b; Werbach, 2018). By voiding the transactions, the core developers made a value judgment in that they favoured the restoration of trust in the community over the ledger’s integrity, understood as the immutable nature of ledger entries. This is because the reversal of the transactions “meant that Ethereum transactions were not truly immune from centralized interference” (Werbach, 2018, p. 68).[5]

Peculiarly, the case of “TheDAO” hack also serves as a curious case highlighting the role that significant miners take in the process of incorporating code amendments, as they independently decide whether or not to follow the core developers’ advice to update their client software. In the case of the TheDAO hack, the community split (DuPont, 2018, 2019). While most miners followed the core developers’ advice, a minority stuck to the old protocol and thereby created an incompatible version of the shared ledger called “Ethereum Classic” (Werbach, 2018). This example demonstrates that only by convincing large proportions of significant miners to adopt the updated implementation developers can turn a software fork into a network fork (Antonopoulos, 2017). Therefore, as a stakeholder group, miners cannot be overruled or circumvented in design decisions regarding protocol changes.[6]

Lastly, the coordinated approach of various cryptocurrency exchanges to delist the Bitcoin-Cash spin-off Bitcoin SV highlights the capability of cryptocurrency exchanges to engage in the negotiation process on how a blockchain-based system should account for human values. Here, two groups of stakeholders proposed different code upgrades for the Bitcoin Cash protocol. One group, surrounding “the developers of the most popular Bitcoin Cash software client, called Bitcoin ABC, proposed a series of upgrades, including smart contract capability.” In contrast, another group, including a mining pool controlling “more than 15 percent of all Bitcoin Cash mining,” proposed a divergent upgrade without such fundamental changes to the system’s capabilities (Orcutt, 2018). The advocates of this alternative upgrade claimed that it adheres more closely to the original ideals of Bitcoin as outlined in early white papers.

Consequently, a *hard fork* occurred, establishing Bitcoin SV as a spin-off of Bitcoin Cash, followed by turmoil within the community and what Orcutt (2019) calls “social media-fueled coin delistings”. Major cryptocurrency exchanges like Kraken or Binance released statements criticising the team behind the newly established Bitcoin SV. KRAKENFX (2019) announced that the behaviour of “the team behind Bitcoin SV” in

[5] Walch (2019b) lists further examples of core developers wielding power in design processes.

[6] Highlighting the power of significant miners in design processes manifested in the discourse on the Bitcoin block-size, which resulted in the hard fork of Bitcoin and Bitcoin Cash in 2017 (DuPont, 2019).

the aftermath of the fork was incongruent with the values held by “Kraken and the wider crypto community”. Binance (2019) questioned whether Bitcoin SV “continues to meet the high level of standard” they expect. Consequently, the two exchanges—among others—stopped exchanging Bitcoin SV on their platforms, which, in turn, lead to “a substantial drop in [Bitcoin SV’s] value” (Orcutt, 2019), restricted access to the system for its users, and diminished economic incentives to participate in the system for Bitcoin SV miners.

4. Findings

Sections 3.1 and 3.2 offer several insights into the challenge of accounting for power in VSD. These allow identifying four crucial factors that co-determine the effects of the distribution of power in socio-technical ecosystems on the applicability of VSD.

Level of decentralisation: Juxtaposing platform-based ecosystems and blockchain-based ecosystems suggests that considering the *level of decentralisation* of the ecosystem is of paramount importance for determining the kind of issues that might occur when accounting for power in the application of VSD. As Shilton and Greene (2019) demonstrate, actors like platform providers at the centre of more centralised ecosystems can assert their ideas of conceptions, weighings, and operationalisations of values to a large extent. Using boundary resources as value levers, they do not just enforce these conceptualisations, weighings, and operationalisations onto the core components of the ecosystem, which they directly control, but also onto the design of peripheral applications.

Conversely, in organisationally more decentralised ecosystems, there is, by definition, no central actor who can similarly assert itself. As shown in the case of open and permissionless blockchains, many actors have the power to impact decisions in the context of how human values are conceptualised, weighed, and operationalised in the design processes. The distribution of power in such ecosystems makes some form of deliberation and coordination inevitable to avoid gridlocks.

Core/periphery: Different issues arise in design decisions concerning an ecosystem’s core components and design decisions concerning peripheral applications. The design of the core components, for the most part, affects more stakeholder groups than the design of peripheral services. Accordingly, negotiation processes are often more complex and conflictual.

Conversely, on the side of peripheral applications, fewer actors are involved. Instead, VSD practitioners have to consider boundary resources that constrain their scope for design. They need to account for the ecosystem's technical and non-technical infrastructure and the actors in charge of it. Section 3.1 underlines the resulting power imbalance in platform-based ecosystems. Platform providers determine the design of boundary resources largely independently and, as a result, determine the scope for design of developers of peripheral services.

Temporality: As demonstrated in the two cases, how actors can exercise power varies considerably. One key difference is *temporality*. Some means of exercising power function *ex-ante*, i.e., actors suppress potential design decisions from being implemented in the first place. Examples of *ex-ante* exercises of power are the design of technical interfaces that predetermine how data can be accessed and managed, how users and peripheral services can interact, and more generally, which criteria peripheral services have to meet in order to be compatible with an ecosystem's technical infrastructure. By utilising technical interfaces, central actors can predetermine how human values like privacy can be conceptualised and operationalised in the entire ecosystem. Other modes of exercising power function *ex-post*, i.e., they interfere with a technical artefact's deployment or usage after an undesired design decision is implemented. Examples of means to exercise power *ex-post* are, for instance, app-store approval processes that central actors can use to exclude specific applications or services from a platform, or the decision of significant miners in blockchain-based ecosystems to omit using a new version of a blockchain protocol after developers deployed it in a software repository.

Phase of VSD: The two cases demonstrate that accounting for power is relevant for making decisions in the conceptual, empirical, and technical *phase of VSD*. They reveal that the way power is situated in the "broader social, economic, cultural, and political networks, institutions, and structures" (Sattarov, 2019, p. 20) of an ecosystem affects how human values are conceptualised (as demonstrated in the analysis of iOS and Android developer forums by Shilton and Greene (2019)), weighed (as demonstrated by Ethereum's core developers' value-judgement leading to recovering the funds after TheDAO hack), operationalised (as demonstrated by Apple's move to change the data interface design for apps to access the MAC addresses of devices), and how the overall process of accounting for values can be evaluated (as illustrated by “social media-fueled coin delistings”). Accounting for power thus concerns VSD practitioners in all phases of VSD.

5. Discussion

These observations allow deriving some points of reference for addressing the challenge of accounting for power. They suggest that the general applicability of VSD and its potential to address power-related issues varies tremendously depending on features of the ecosystem and the role that a given artefact is supposed to play in it. While the distribution of power within an ecosystem, in some cases, hinders the application of VSD, it appears to accommodate the approach in other cases. Though this is true for both more centralised and more decentralised systems, the decisive factors differ.

In more centralised ecosystems like platform-based ecosystems, boundary resources function as obligatory passage points (Law & Callon, 1992; see also Callon, 1984) for peripheral applications and constrain the application’s design process. These constraints cut back on the agency of developers and can either obstruct or compel design decisions that promote or demote the realisation of specific values.[7] Shilton and Greene (2019) outline discussions from iOS and Android developer forums that illustrate this lack of agency of developers of peripheral applications in platform-based ecosystems. Here, developers are often required to interpret and realise a concept of privacy that is predefined and manifests, e.g., in the platform’s boundary resource design. As (Greene & Shilton, 2017, p. 16) note, “platforms govern design by promoting particular ways of doing privacy, training devs on those practices, and (to varying degrees) rewarding or punishing them based on their performance”. Thus, while developers of peripheral applications can always decide to collect less data, privacy here, for the most part, is whatever the platform providers define as privacy. A meaningful application of VSD is virtually non-viable for developers of peripheral applications in such settings.

[7] This issue only comes into play where design decisions or other actions by platform providers are in conflict with (the operationalisation of) a value that VSD practitioners aim to account for. Applying VSD to account for other values is still possible for developers of peripheral applications.

While Hestres (2013) and van van Dijck et al. (2018) outline how concerted efforts of various stakeholder groups can principally have success in appealing to platform providers and lead to changes in boundary resource design, this commonly is not part of the design process of individual technical artefacts and thus outside the scope of VSD. Nevertheless, grassroots efforts proved effective in many cases and should be considered a tool to create the necessary environment for an application of VSD by VSD practitioners in more centralised socio-technical ecosystems.[8]

However, if Apple's boundary resource design is assessed regarding its direct impact on the realisation of values, it's apparent that it gives users more autonomy by allowing them to configure the location data usage (Apple, 2019) and to protect their privacy by eradicating access to a device's MAC address (Butts, 2017). Thus, this case shows that powerful actors such as platform providers can also encourage "ethical practice within their ecosystems" (Shilton & Greene, 2019, p. 144) by making use of a carefully considered boundary resource design. Regarding privacy, Shilton and Greene (2016, n.p.) describe this phenomenon as "a 'trickledown privacy' effect in which platform providers exercise strong power over privacy definitions". As platform providers can shape the conceptualisation of values within the ecosystem more generally, similar effects can be realised with other values (Shilton & Greene, 2019).[9] Therefore, if VSD is used in boundary resource design, it enables the *value sensitive shaping of ecosystems*.

While the means by which the platform providers exert power in the discussed examples are primarily technical on the surface, they also affect developers of peripheral applications economically. For instance, as Ausloos and Veale (2020, p. 138) note, platform providers can utilise a restrictive API design to "break an entire set of business models" that rely on specific data streams through the respective APIs (see also Bucher, 2013; Leerssen et al., 2019). Thus, platform providers can use technical means such as API design to exert economic pressure on other actors within the respective ecosystem by affecting the economic viability and potential profitability

[8] For a current example, observe the current dispute between Apple and the Coalition for App Fairness (Gartenberg, 2020).

[9] When using VSD in the design of individual technical artefacts, platform providers might have to deal with value conflicts during attempts to engage in the value-sensitive shaping of an ecosystem. These conflicts may arise between two or more human values or between a human value—such as privacy—and instrumental values—such as cost-efficiency or usability.

of business models behind peripheral applications. Such strategic exploitation of the API design as a tool “to exclude certain business or functionality from integration” (Ausloos & Veale, 2020, p. 138) can establish economic constraints on the scope for design of VSD practitioners at the periphery of ecosystems.

In more decentralised ecosystems, the challenge of accounting for power manifests differently. Here, the negotiation processes among the involved actors on how to account for human values can lead to gridlock. This is because various actors with divergent incentives and interests may disagree as to how human values are conceptualised, weighed, or operationalised. Applying VSD in such cases could ensure that various stakeholder groups are represented in decision-making processes and balance the interests of different stakeholder groups without calling into question the ecosystem's decentralised nature. However, in more decentralised ecosystems, VSD practitioners need to ensure that actors who are not directly involved in design decisions, but affected by them, are also taken into account.

These differences in ecosystems suggest that if there is freedom of choice, the selection of the ecosystem that VSD practitioners embed an artefact in has a significant effect on how human values can be accounted for in the artefact’s design. For instance, Atzori and Ulieru (2017) argue that research on platformisation, i.e., “the penetration of the infrastructures, economic processes, and governmental frameworks of platforms in different economic sectors and spheres of life” (Poell et al., 2019, pp. 5–6) is indicating that “the concept of distributive justice / distributive efficiency [is] strongly dependent on platform architectural design and they are unlikely to be achieved in centralized, two-sided markets” (Atzori & Ulieru, 2017, pp. 4–5). However, due to the quasi-monopolistic position of many platforms (Eaton et al., 2015), such a choice does not always exist for developers of peripheral applications if they want to attract a larger user group. Furthermore, that developers of peripheral applications need to consider not only a platform’s current features, but also the platform provider’s means (technical, economic, or other) to exert power in the future, further complicates the selection process.

Moreover, as the means of different actors to exercise power concern various phases of the design process and can come into play even after deployment, VSD practitioners have to consider the matter continuously: from early conceptualisations of values to the process of operationalising and implementing values to the deployment of artefacts and the evaluation of the design decisions related to values later on. More specifically, since both the development of core components of an ecosystem and peripheral applications often continue after deployment in the form of a constant redesign (Eaton et al., 2015), VSD similarly has to incorporate continuous monitoring of the respective ecosystem's modifications, updates, developments and related effects on the distribution of power within the ecosystem. While the foundational texts of VSD in principal already set out the approach as extending over all of these phases (Friedman et al., 2008; Friedman & Hendry, 2019), in practice, most practitioners do not perform several iterations of the three phases over the entire length of the design process (Winkler & Spiekermann, 2018). Therefore, it is essential to stress the importance of a continuous application of VSD once more.

Furthermore, the findings of this paper suggest that the spectrum of tasks involved in VSD expands if applied in the design of artefacts embedded in vast and complex socio-technical ecosystems. Some tasks, such as monitoring the distribution of power in the ecosystem over extended periods of time or dealing with platform monopolies, appear to be too extensive to be addressed by individual VSD practitioners or even small development teams. Therefore, the range of tasks needs to be distributed over more actors if they are to remain manageable. Regulatory authorities, in particular, must play a role in addressing some of these challenges. In particular, challenges arising due to 1) (quasi-) monopolistic players, 2) the complexity of continuously monitoring the manifold actor constellations and distribution of power within an ecosystem, and 3) boundary resource design that prevents developers of peripheral applications from accounting for values surpass the capabilities of VSD practitioners and the scope of VSD in a traditional sense.

Dealing with (quasi-) monopolistic actors, especially platform providers, is in the domain of antitrust and competition authorities (Crémer et al., 2019; Kommission Wettbewerbsrecht 4.0, 2019; Monopolkommission, 2015), which therefore play a crucial role in ensuring the applicability of VSD. Furthermore, the European Commission's recent proposal for the Digital Markets Act (European Commission, 2020c) contains several propositions for concrete regulatory measures aiming to curb the quasi-monopolistic standing of many platform providers. Relevant here are, for instance, the proposed requirements for gatekeepers to "allow business partners," such as the providers of peripheral applications in a platform-based ecosystem, "to offer the same products or services to end users through third party online intermediation services at prices or conditions that are different from those offered through the online intermediation services of the gatekeeper"[10] (European Commission, 2020c, art. 5 (b)) or to "provide effective portability of data generated through the activity of a business user or end user [...]" (European Commission, 2020c, art. 6 (h)). Thereby, regulation built on the European Commission's proposal for a Digital Markets Act could help to counteract "winner-takes-all dynamics" (Anderson & Mariniello, 2021) that favour the development of (quasi-)monopolies. In turn, such a development would provide more choices to select a suitable platform (or suitable platforms) for developers of peripheral applications and make it a more viable option to integrate this selection process in the application of VSD.

Additionally, establishing oversight institutions like the recently launched AI Observatory of the German Federal Ministry of Labour and Social Affairs (Bundesregierung, 2020) or a competence centre for algorithmic systems, as proposed by the German Data Ethics Commission (Datenethikkommission, 2019), could play a crucial role in the monitoring of platform-based ecosystems. The proposal for the Digital Markets Act outlines further supportive measures. Especially the requirement for gatekeepers "to refrain from preventing or restricting business users from raising issues with any relevant public authority relating to any practice of gatekeepers" (European Commission, 2020c, art. 5 (d)) is crucial here, as it would allow for a closer collaboration of regulatory authorities and developers of peripheral applications. If cooperating closely, oversight institutions and developers of peripheral applications could identify the most problematic practices of platform providers jointly in a bottom-up approach and lay the groundwork for possible future regulation and governance that addresses the most urgent issues for VSD practitioners.

Furthermore, in future regulatory frameworks, regulatory authorities should consider an ecosystem's boundary resource composition when determining. As outlined above, iOS is a closed platform. Apple controls and governs the unique distribution channel and, thus, establishes itself as an obligatory passage point. Developers of peripheral applications need to follow Apple's guidelines closely since there is no alternative distribution channel for applications on iOS devices. This is a conscious decision by

[10] Note that the term "gatekeeper" is defined more narrowly by the European Commission than the term "platform provider" that is used in this paper (see European Commission, 2020c).

Apple which brings the company a wide range of business benefits. Yet, from a regulatory perspective, this decision could also be linked to stricter obligations for Apple since it constrains the scope for design of developers of peripheral applications and predetermines to what degree they can account for human values in design decisions. If platform providers use their power to exert influence over design decisions and limit access to alternative distribution channels for developers of peripheral applications, it seems reasonable to link these activities to a stricter regime of obligations. Suppose platform providers engage in exercising control over how developers of peripheral applications account for human values in the technical design of their applications and shape the ecosystem more proactively. In that case, this *value sensitive shaping* of the ecosystem should be subject to increased scrutiny by regulators. Conversely, if platforms refrain from exercising control over how developers of peripheral applications account for human values in the technical design of their applications, the focus of regulators should shift more to the actors at the ecosystem's periphery.

6. Conclusion

Technical design in accordance with human values is increasingly considered a building block for shaping the digital future. VSD is a long-standing and well-established approach for achieving design in accordance with human values. This paper shows that the integration of technical artefacts in increasingly vast and complex socio-technical ecosystems with power distributed over various actors affects the applicability of VSD in multiple ways. Several factors determine how this challenge manifests in practice. This paper identifies 1) the *level of decentralisation* of the ecosystem in question, 2) whether VSD is applied regarding the design of components of an ecosystem's *core* or *periphery*, 3) the *temporality* of the exercise of power, and 4) the *phase of VSD* in which power is exercised in as four of these factors.

Adopting a power-sensitive ecosystem perspective provides some reference points for addressing the challenge of accounting for power. While in some constellations, the application of VSD appears to be less applicable since the scope for design of developers is restricted, other constellations appear to accommodate the approach. On the one hand, these are cases where a multitude of assertive actors engage in decision-making processes regarding specific design choices that result in conflicts or even gridlock. Here, VSD can provide a structured approach that supports resolving these conflicts and balances the interests of different stakeholder groups. VSD also appears to be of particular importance for the design of an ecosystem's core components, such as boundary resources. Here, individual design decisions can shape entire ecosystems in accordance with human values (see Shilton, 2012). For this reason, highly centralised ecosystems are also potentially more attractive to regulatory authorities because important nodes and actors in such ecosystems are more easily identifiable and addressable.

Dealing with (quasi-)monopolistic players, accounting for the complexity of continuously monitoring the complex actor constellations and distribution of power within an ecosystem, and addressing boundary resource design preventing developers of peripheral applications from accounting for values emerge as significant novel challenges when applying VSD in vast and complex socio-technical ecosystems. However, recent proposals for establishing new oversight institutions (Bundesregierung, 2020; Datenethikkommission, 2019) and new regulatory approaches such as the Digital Markets Act and the Digital Services Act (European Commission, 2020b, 2020c) indicate that regulatory authorities can support VSD practitioners in overcoming these challenges. Furthermore, in the future, close cooperation between oversight institutions and VSD practitioners can reveal problematic practices of powerful actors in socio-technical ecosystems and thereby lay the foundation for further regulatory action.

Lastly, a lesson that can be drawn for further research in the field of VSD is that developing a unified framework for dealing with power imbalances between stakeholders in socio-technical ecosystems does not seem to be an attainable goal because the way that power manifests in different ecosystems varies substantively. Thus, instead of aiming for a unified framework, practitioners need to make calls on adequate procedures on a case-by-case basis. Future research, therefore, should aim at advancing the understanding of the actor constellations in socio-technical ecosystems and the distribution of power within them. In particular, in-depth comparative analyses of various socio-technical ecosystems, the distribution of power among the actors involved in them, and the human values expressed in the design of their boundary resource design could provide valuable and more readily applicable insights for VSD practitioners.

Acknowledgements

Mattis Jacobs, Christian Kurtz, Judith Simon, and Tilo Böhmann acknowledge financial support from the Hamburg Ministry of Science, Research and Equality in the project Information Governance Technologies—Ethics, Policies, Architectures, Engineering—with reference LFF-FV 34. Mattis Jacobs and Judith Simon acknowledge financial support from the German Federal Ministry of Education and Research in the project GOAL—about algorithmic behaviour control and artificial intelligence—with reference 01IS19020.

Publication's References

Alsindi, W. Z., & Lotti, L. (2021). Mining. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1551

Alstyne, M., Parker, G., & Choudary, S. P. (2016). Pipelines, platforms, and the new rules of strategy. *Harvard Business Review*, *94*(4), 54–62.

Anderson, J., & Mariniello, M. (2021). Regulating big tech: The Digital Markets Act [Blog post]. *Bruegel*. https://www.bruegel.org/2021/02/regulating-big-tech-the-digital-markets-act/

Antonopoulos, A. (2014). Bitcoin security model: Trust by computation. *O'Reilly Radar*. http://radar.oreilly.com/2014/02/bitcoin-security-model-trust-by-computation.html

Antonopoulos, A. (2017). *Mastering Bitcoin: Programming the open Blockchain* (2nd ed.). O'Reilly.

Apple. (2019). *Turn Location Services and GPS on or off on your iPhone*. Apple Support. https://support.apple.com/en-us/HT207092

Apple. (2020). *App Tracking Transparency: Request user authorization to access app-related data for tracking the user or the device*. Apple Developer Documentation. https://developer.apple.com/documentation/apptrackingtransparency

Atzori, M., & Ulieru, M. (2017). Architecting the eSociety on Blockchain: A Provocation to Human Nature. *SSRN Electronic Journal. Advance Online Publication*. https://doi.org/10.2139/ssrn.2999715

Ausloos, J., & Veale, M. (2020). *Researching with Data Rights*.

Baldwin, C., & Woodard, C. J. (2009). The Architecture of Platforms: A Unified View. In A. Gawer (Ed.), *Platforms, Markets and Innovation* (pp. 19–46). Edward Elgar.

Becker, M., & Bodó, B. (2021). Trust in blockchain-based systems. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1555

Binance. (2019, April 15). *Binance Will Delist BCHSV*. Binance Latest News. https://binance.zendesk.com/hc/en-us/articles/360026666152

Bodó, B., Brekke, J. K., & Hoepman, J. -H. (2021). Decentralisation: A multidisciplinary perspective. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1563

Borning, A., Waddell, P., & Förster, R. (2008). Urbansim: Using Simulation to Inform Public Deliberation and Decision-Making. In R. Sharda, S. Voß, H. Chen, L. Brandt, V. Gregg, R. Traunmüller, S. Dawes, E. Hovy, A. Macintosh, & C. A. Larson (Eds.), *Integrated Series In Information Systems. Digital Government* (Vol. 17, pp. 439–464). Springer US. https://doi.org/10.1007/978-0-387-71611-4_22

Botsman, R. (2017). *Who can you trust? How technology brought us together and why it might drive us apart* (1st ed.). Public Affairs.

Brekke, J. K., & Alsindi, W. Z. (2021). Cryptoeconomics. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1553

Brey, P. (2008). The Technological Construction of Social Power. *Social Epistemology*, *22*(1), 71–95. https://doi.org/10.1080/02691720701773551

Brey, P. (2010). Values in technology and disclosive computer ethics. In L. Floridi (Ed.), *The Cambridge handbook of information and computer ethics* (pp. 41–58). Cambridge Univerity Press. https://doi.org/10.1017/CBO9780511845239.004

Bucher, T. (2013). Objects of intense feeling: The case of the Twitter API. *Computational Culture*, *3*.

Bundesregierung. (2020). *Technologie soll dem Menschen dienen: KI-Observatorium nimmt Arbeit auf* [Press release]. Presse- und Informationsamt der Bundesregierung. https://www.bundesregierung.de/breg-de/aktuelles/ki-oberservatorium-1726794

Butts, J. (2017). Thanks to Misuse, Apps Can't View Mac Addresses on iOS 11. *Mac Observer*. https://www.macobserver.com/news/product-news/apps-cant-view-mac-addresses-on-ios-11/

Callon, M. (1984). Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review*, *32*(1_suppl), 196–233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x

Constantinides, P., Henfridsson, O., & Parker, G. G. (2018). Introduction—Platforms and Infrastructures in the Digital Age. *Information Systems Research*, *29*(2), 381–400. https://doi.org/10.1287/isre.2018.0794

Crémer, J., Montjoye, Y.-A., & Schwitzer, H. (2019). *Competition Policy for the Digital Era* (Report KD-04-19-345-EN-N). Publications Office of the European Union. http://doi.org/10.2763/407537

Czeskis, A., Dermendjieva, I., Yapit, H., Borning, A., Friedman, B., Gill, B., & Kohno, T. (2010). Parenting from the pocket. In L. F. Cranor (Ed.), *Proceedings of the Sixth Symposium on Usable Privacy and Security—SOUPS '10* (p. 1). https://doi.org/10.1145/1837110.1837130

Datenethikkommission. (2019). *Gutachten der Datenethikkommission* [Report]. Bundesministerium des Innern, für Bau und Heimat. https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf

De Filippi, P. (2017). In Blockchain we Trust: Vertrauenslose Technologie für eine vertrauenslose Gesellschaft. In Rudolf-Augstein-Stiftung (Ed.), *Reclaim Autonomy: Selbstermächtigung in der digitalen Weltordnung* (edition suhrkamp, Vol. 2714, pp. 53–81).

De Filippi, P., Mannan, M., & Reijers, W. (2020). Blockchain as a confidence machine: The problem of trust & challenges of governance. *Technology in Society*, *62*. https://doi.org/10.1016/j.techsoc.2020.101284

De Filippi, P., & Wright, A. (2018). *Blockchain and the law: The rule of code*. Harvard University Press.

DuPont, Q. (2018). Experiments in Algorithmic Governance: A history and ethnography of 'The DAO,' a failed Decentralized Autonomous Organization. In M. Campbell-Verduyn (Ed.), *Bitcoin and beyond: Cryptocurrencies, blockchains and global governance*. Routledge. https://doi.org/10.4324/9781315211909-8

DuPont, Q. (2019). *Cryptocurrencies and blockchains*. John Wiley & Sons.

DuPont, Q., & Maurer, B. (2015, June 23). Ledgers and Law in the Blockchain. *Kings Review*. https://www.kingsreview.co.uk/essays/ledgers-and-law-in-the-blockchain

Eaton, B., Elaluf-Calderwood, S., Sørensen, C., & Yoo, Y. (2015). Distributed Tuning of Boundary Resources: The Case of Apple's iOS Service System. *MIS Quarterly*, *39*(1), 217–243. https://doi.org/10.25300/MISQ/2015/39.1.10

European Commission. (2019). *Building Trust in Human-Centric Artificial Intelligence*. https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

European Commission. (2020). *On Artificial Intelligence—A European approach to excellence and trust* (White Paper COM(2020) 65 final). https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive, COM(2020) 825 final (2020). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en

Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on contestable and fair markets in the digital sector (Digital Markets Act), COM(2020) 842 final (2020). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0842&from=en

Feiler, P., Sullivan, K., Wallnau, K., Gabriel, R., Goodenough, J., Linger, R., Longstaff, T., Kazman, R., Klein, M., Northrop, L., & Schmidt, D. (2006). *Ultra-Large-Scale Systems: The Software Challenge of the Future*. Software Engineering Institute, Carnegie Mellon University.

Friedman, B., & Hendry, D. G. (2019). *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press.

Friedman, B., Hendry, D. G., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction*, *11*(2), 63–125. https://doi.org/10.1561/1100000015

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value Sensitive Design and Information Systems. In K. E. Himma & H. T. Tavani (Eds.), *The Handbook of Information and Computer Ethics* (pp. 69–101). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470281819.ch4

Friedman, B., Kahn, P., Hagman, J., Severson, R., & Gill, B. (2006). The Watcher and the Watched: Social Judgments About Privacy in a Public Place. *Human-Computer Interaction*, *21*(2), 235–272. https://doi.org/10.1207/s15327051hci2102_3

Gartenberg, C. (2020, September 24). Spotify, Epic, Tile, Match, and more are rallying developers against Apple's App Store policies: As the 'Coalition for App Fairness'. *The Verge*. https://www.theverge.com/2020/9/24/21453745/spotify-epic-tile-match-coalition-for-app-fairness-apple-app-store-policies-protest

Ghazawneh, A., & Henfridsson, O. (2013). Balancing platform control and external contribution in third-party development: The boundary resources model. *Information Systems Journal*, *23*(2), 173–192. https://doi.org/10.1111/j.1365-2575.2012.00406.x

Giungato, P., Rana, R., Tarabella, A., & Tricase, C. (2017). Current Trends in Sustainability of Bitcoins and Related Blockchain Technology. *Sustainability*, *9*(12), 2214. https://doi.org/10.3390/su9122214

Greene, D., & Shilton, K. (2018). Platform Privacies: Governance, Collaboration, and the Different Meanings of "Privacy" in iOS and Android Development. *New Media & Society*, *20*(4), 1640–1657. https://doi.org/10.1177/1461444817702397

Haugaard, M. (2010). Power: A 'family resemblance' concept. *European Journal of Cultural Studies*, *13*(4), 419–438. https://doi.org/10.1177/1367549410377152

Hawlitschek, F., Notheisen, B., & Teubner, T. (2018). The limits of trust-free systems: A literature review on blockchain technology and trust in the sharing economy. *Electronic Commerce Research and Applications*, *29*, 50–63. https://doi.org/10.1016/j.elerap.2018.03.005

Hestres, L. (2013). App neutrality: Apple's app store and freedom of expression online. *International Journal of Communication*, *7*, 1265–1280. https://ijoc.org/index.php/ijoc/article/view/1904

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI* [Report]. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Introna, L. D. (2005). Disclosive Ethics and Information Technology: Disclosing Facial Recognition Systems. *Ethics and Information Technology*, *7*(2), 75–86. https://doi.org/10.1007/s10676-005-4583-2

Jacobs, M. (2020). How Implicit Assumptions on the Nature of Trust Shape the Understanding of the Blockchain Technology. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00410-x

Karhu, K., Gustafsson, R., & Lyytinen, K. (2018). Exploiting and defending open digital platforms with boundary resources: Android's five platform forks. *Information Systems Research : ISR : An Information Systems Journal of the Institute for Operations Research and the Management Sciences*, *29*(2), 479–497.

Kommission Wettbewerbsrecht. (2019). *A New Competition Framework for the Digital Economy: Report by the Commission 'Competition Law 4.0'* [Report]. https://www.bmwi.de/Redaktion/EN/Downloads/a/a-new-competition-framework.pdf?__blob=publicationFile&v=2

KRAKENFX. (2019). Kraken is Delisting BSV [Blog post]. *Kraken*. https://blog.kraken.com/post/2274/kraken-is-delisting-bsv/

Law, J., & Callon, M. (1992). The Life and Death of an Aircraft: A Network Analysis of Technical Change. In W. E. Bijker & J. Law (Eds.), *Inside technology. Shaping technology/building society: Studies in sociotechnical change* (pp. 21–52). MIT Press.

Leerssen, P., Ausloos, J., Zarouali, B., Helberger, N., & Vreese, C. H. (2019). Platform Ad Archives: Promises and Pitfalls. *Internet Policy Review*, *8*(4), 1–21. https://doi.org/10.14763/2019.4.1421

Leswing, K. (2021). *Apple exec warns it may remove apps that track users without permission*. https://www.cnbc.com/2020/12/08/apple-may-remove-apps-that-track-users-without-permission-in-2021.html

Lusch, R. F., & Nambisan, S. (2015). Service Innovation: A Service-Dominant Logic Perspective. *MIS Quarterly*, *39*(1), 155–175. https://doi.org/10.25300/MISQ/2015/39.1.07

McConahy, A., Eisenbraun, B., Howison, J., Herbsleb, J. D., & Sliz, P. (2012). *Techniques for monitoring runtime architectures of socio-technical ecosystems*. Workshop on Data-Intensive Collaboration in Science and Engineering (CSCW 2012).

Monopolkommission. (2015). *Competition policy: The challenge of digital markets. Special Report by the Monopolies Commission pursuant to section 44(1)(4) of the Act Against Restraints on Competition* (Special Report No. 68). Monopolies Commission. https://www.monopolkommission.de/images/PDF/SG/s68_fulltext_eng.pdf

Morabito, V. (2017). *Business Innovation Through Blockchain: The B3 Perspective*. Springer International Publishing. https://doi.org/10.1007/978-3-319-48478-5

Mueller, M., & Heger, O. (2018). *Health at any Cost? Investigating Ethical Dimensions and Potential Conflicts of an Ambulatory Therapeutic Assistance System through Value Sensitive Design*. 39th International Conference on Information Systems (ICIS). https://aisel.aisnet.org/icis2018/healthcare/Presentations/17/

Nathan, L. P. (2012). Sustainable information practice: An ethnographic investigation. *Journal of the American Society for Information Science and Technology*, *63*(11), 2254–2268. https://doi.org/10.1002/asi.22726

Nathan, L. P., Klasnja, P. V., & Friedman, B. (2007). Value scenarios: A technique for envisioning systemic effects of new technologies. *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, 2585–2590. https://doi.org/10.1145/1240866.1241046

Orcutt, M. (2018). *Chain Letter #102: You can go your own way*. https://mailchi.mp/technologyreview/chain-letter-767541?e=93dc606e34&utm_campaign=chain_letter.unpaid.engagement&utm_source=hs_email&utm_medium=email&utm_content=71892724&_hsenc=p2ANqtz--2S3Tqvwcm1BQc_Eb-ArlTKDA-f9cAdBdc6Lxq67nmBy3Y2Z48OyMOEMW__mczpT4YRw2w-7sUtAvryWgLnYMUo_VkOxYs6DqYzXyno1pRK7AGncM&_hsmi=71892724

Orcutt, M. (2019). Chain Letter #139: On social media-fueled coin delistings [Blog post]. *Cryptocurrency News Now!* https://cryptocurrency-news-now.blogspot.com/2019/04/139-on-social-media-fueled-coin.html

Perez, S. (2020). Coalition for App Fairness, a group fighting for app store reforms, adds 20 new partners. *TechCrunch*. https://techcrunch.com/2020/10/21/coalition-for-app-fairness-a-group-fighting-for-app-store-reforms-adds-20-new-partners/

Poell, T., Nieborg, D., & van Dijck, J. (2019). Platformisation. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1425

Reuver, M. de, Sørensen, C., & Basole, R. C. (2018). The Digital Platform: A Research Agenda. *Journal of Information Technology*, *33*(2), 124–135. https://doi.org/10.1057/s41265-016-0033-3

Sattarov, F. (2019). *Power and technology: A philosophical and ethical analysis*. Rowman et Littlefield.

Shilton, K. (2012). Values Levers. *Science, Technology, & Human Values*, *38*(3), 374–397. https://doi.org/10.1177/0162243912436985

Shilton, K., & Greene, D. (2016, March 15). *Because privacy: Defining and legitimating privacy in ios development*. iConference 2016. https://doi.org/10.9776/16229

Shilton, K., & Greene, D. (2019). Linking Platforms, Practices, and Developer Ethics: Levers for Privacy Discourse in Mobile Application Development. *Journal of Business Ethics*, *155*(1), 131–146. https://doi.org/10.1007/s10551-017-3504-8

Simon, J. (2016). Values in Design. In J. Heesen (Ed.), *Handbuch Medien- und Informationsethik* (pp. 357–364). J.B. Metzler.

Simon, J., Wong, P. -H., & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1534

Swan, M. (2015). *Blockchain: Blueprint for a new economy* (First edition.). O'Reilly.

Tiwana, A., & Konsynski, B. (2010). Complementarities Between Organizational IT Architecture and Governance Structure. *Information Systems Research*, *21*(2), 288–304. https://doi.org/10.1287/isre.1080.0206

Tiwana, A., Konsynski, B., & Bush, A. A. (2010). Research Commentary—Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics. *Information Systems Research*, *21*(4), 675–687. https://doi.org/10.1287/isre.1100.0323

van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

van House, N. A. (2004). Science and technology studies and information studies. *Annual Review of Information Science and Technology*, *38*, 3–86. https://doi.org/10.1002/aris.1440380102

Waddell, P., Wang, L., & Liu, X. (2008). UrbanSim: An evolving planning support system for evolving communities. In R. K. Brail (Ed.), *Planning Support Systems for Cities and Regions* (pp. 103–138). Lincoln Institute for Land Policy.

Walch, A. (2019). In Code(rs) We Trust: Software Developers as Fiduciaries in Public Blockchains. In P. Hacker, I. Lianos, G. Dimitropoulos, & S. Eich (Eds.), *Regulating Blockchain: Techno-Social and Legal Challenges* (pp. 58–82). Oxford University Press. https://doi.org/10.1093/oso/9780198842187.003.0004

Walton, R., & DeRenzi, B. (2009). Value-Sensitive Design and Health Care in Africa. *IEEE Transactions on Professional Communication*, *52*(4), 346–358. https://doi.org/10.1109/TPC.2009.2034075

Warnier, M., Dechesne, F., & Brazier, F. (2015). Design for the Value of Privacy. In J. Hoven, V. P. E., & I. van de Poel (Eds.), *Handbook of ethics, values, and technological design: Sources, theory, values and application domains* (pp. 431–445). Springer. https://doi.org/10.1007/978-94-007-6970-0_17

Weber, M. (2019). *Economy and Society: A New Translation*. Harvard University Press.

Werbach, K. (2018). *The Blockchain and the New Architecture of Trust*. MIT Press.

Winkler, T., & Spiekermann, S. (2018). Twenty years of value sensitive design: A review of methodological practices in VSD projects. *Ethics and Information Technology*, *18*(4), 185. https://doi.org/10.1007/s10676-018-9476-2

Wright, A., & De Filippi, P. (2015). *Decentralized blockchain technology and the rise of lex cryptographia*. https://doi.org/10.2139/ssrn.2580664

Yoo, D., Derthick, K., Ghassemian, S., Hakizimana, J., Gill, B., & Friedman, B. (2016). Multi-lifespan Design Thinking. In J. Kaye, A. Druin, C. Lampe, D. Morris, & J. P. Hourcade (Eds.), *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4423–4434). https://doi.org/10.1145/2858036.2858366

9 Publication 3:

Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed By the European Commission

Authored by Mattis Jacobs and Judith Simon

Jacobs, M., & Simon, J. (2021). Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed by the European Commission. *Digital Society, 1*(1), Article 6. https://doi.org/10.1007/s44206-022-00009-z

Citation style and bibliography harmonized

Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed By the European Commission

Mattis Jacobs, Judith Simon

Abstract

The emergence and increasing prevalence of Artificial Intelligence (AI) systems in a growing number of application areas brings about opportunities but also risks for individuals and society as a whole. To minimize the risks associated with AI systems and to mitigate potential harm caused by them, recent policy papers and regulatory proposals discuss obliging developers, deployers, and operators of these systems to avoid certain types of use and features in their design. However, most AI systems are complex socio-technical systems in which control over the system is extensively distributed. In many cases, a multitude of different actors is involved in the purpose setting, data management and data preparation, model development, as well as deployment, use, and refinement of such systems. Therefore, determining sensible addressees for the respective obligations is all but trivial. This article discusses two frameworks for assigning obligations that have been proposed in the European Commission's whitepaper *On Artificial Intelligence—A European approach to excellence and trust* and the proposal for the Artificial Intelligence Act respectively. The focus is on whether the frameworks adequately account for the complex constellations of actors that are present in many AI systems and how the various tasks in the process of developing, deploying, and using AI systems, in which threats can arise, are distributed among these actors.

1. Introduction

The emergence and increasing prevalence of AI (Artificial Intelligence) systems in a growing number of application areas brings about opportunities but also risks for individuals and society as a whole. To minimize the risks associated with AI systems and to mitigate potential harm caused by them, recent policy papers and regulatory proposals discuss obliging developers, deployers, and operators of these systems to avoid certain types of use and features in their design that bring about "risks or negative consequences for individuals or the society" (European Commission, 2021c, p. 1) by threatening the realization of ethical values, the consideration of ethical principles, and fundamental rights (Datenethikkommission, 2019; European Commission, 2019, 2020a, 2021c; HLEG-AI, 2019).

However, most AI systems are complex socio-technical systems in which control over the system is extensively distributed. In many cases, a multitude of different actors is involved in the purpose setting, data management and data preparation, model development, as well as deployment, use, and refinement of such systems. And, as Barocas and Selbst (2016), Danks and London (2017), and others demonstrate, threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights can manifest during all these tasks. Therefore, determining sensible addressees for the respective obligations is all but trivial.

This article discusses two frameworks for assigning obligations that have been proposed in the European Commission's (EC) 2020 whitepaper *On Artificial Intelligence—A European approach to excellence and trust* (European Commission, 2020a) and the EC's proposal for the Artificial Intelligence Act (AI Act) (European Commission, 2021c) respectively. The EC's whitepaper *On Artificial Intelligence* proposes a capability-based approach for assigning obligations arguing that "the actor(s) who is (are) best placed to address" the respective issue should be obliged to do so (European Commission, 2020a). On the contrary, the AI Act argues that the "majority of all obligations" should fall on the person or body "placing [the AI system] on the market or putting it into service under its own name or trademark" (Veale & Zuiderveen Borgesius, 2021) and thus focuses on rather fixed addressees.

While the two proposals argue that their respective framework for assigning obligations is appropriate (European Commission, 2020a, p. 22, 2021c, p. 31), neither of them engages in a comparative analysis or in-depth discussion of both frameworks and their respective advantages and disadvantages. Therefore, the rationale behind the shift from the capability-based reasoning of the EC's whitepaper *On Artificial*

Intelligence to the reasoning based on fixed addressees in the AI Act is neither readily evident, nor does it follow from one of the proposals.[1]

Therefore, this article attempts to evaluate both frameworks to assess if the EC's shift from one proposal to the other is normatively appropriate. The focus is on whether the respective frameworks adequately account for the complex constellations of actors that are present in many AI systems and how the various tasks in the process of developing, deploying, and operating AI systems, in which threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights can arise, are distributed among these actors.

To do so, Sect. 2 provides an overview of the different tasks that exist in the process of developing, deploying, and operating AI systems and the actors involved in performing these tasks. Section 3 sets out the two frameworks proposed by the EC for assigning obligations to these actors in more detail. Section 4 links the threats posed by AI systems to the various tasks in the process of developing, deploying, and operating AI systems. Based on these analyses, Sect. 5 discusses the merit of the shift from the capability-based framework outlined in the EC's whitepaper *On Artificial Intelligence* to the framework based on fixed addressees outlined in the AI Act. Section 6 concludes by summarizing the article's results and outlining the unresolved challenges. Furthermore, it sets out how further regulation and future research can support addressing these challenges.

2. AI Systems as Complex Socio-technical Systems

The EC defines AI systems in a very broad sense with a relative openness regarding the technical approach of the system and its application context (European Commission, 2020a, p. 2, 2021a, p. 1, 2021c, p. 39). However, most of the threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights as well as technical features discussed in current policy papers and regulatory proposals concern machine learning-based AI systems with a narrow scope

[1] The AI Act explicitly builds on the *whitepaper On Artificial Intelligence* in several passages (European Commission, 2021c, 1, 5, 7-9). However, it does not address the different frameworks regarding assigning obligations.

of application.[2] Here, especially systems that make decisions or provide the basis for decisions (for instance, in the form of predictions or recommendations) that concern individuals or groups are regarded as ethically relevant.[3]

AI systems based on machine learning that make decisions or provide the basis for decisions consist of several components. As Krafft et al. (2020) note, especially algorithms of two types are involved: one algorithm that infers "decision rules from data" and another one that "merely uses these decision rules to score or classify cases." The algorithm of the first type, "the learning method," and "the decision rules generated from it" constitute the core of such ADM systems, whereas the "scoring or classification algorithm, in contrast, is usually rather simple as it merely applies the trained statistical model" (Krafft et al., 2020, p. 121).

To "learn from experience" (Russell & Norvig, 1995, p. 518) requires capabilities to "collect, store, and process digital data [...] and to utilize vast data sets to train and feed machine learning algorithms that rely upon feedback loops to improve their own performance" (Yeung, 2019, p. 21). Indeed, most of the recent resurge in AI is based not on the novelty of theoretical models—many of which were "theorized and developed decades [...] ago"—but on the availability of data and the capability to handle and process it at scale (Keller et al., 2018, pp. 7–8). For instance, as Quan and Sanderson (2018) note, natural language processing (NLP) would "not be possible without millions of human speech samplings, recorded and broken down," provided as training data. The capabilities to handle and process data as necessary became available, not least due to the proliferation of more potent hardware. Keller et al. (2018, pp. 7–8) elaborate that "storage technology is now mature enough to store and shift vast amounts of training data [and that] the development of GPUs for graphics and gaming applications have made massive parallelized computing significantly cheaper than when neural networks were invented."

[2] Such AI systems based on machine learning need to be distinguished from knowledge-based "expert systems," in which problem-specific knowledge of experts is formalized, allowing to automate rule-based decisions "on narrowly defined tasks" (Russell & Norvig, 1995, p. 255). The statements made in this article about the constellation of actors involved in the development, deployment, and use of AI systems based on machine learning do not necessarily apply to expert systems.

[3] Many arguments and direct quotations in this article refer to these systems exclusively and therefore use the term "algorithmic decision-making system" or "ADM system" instead of "AI system.".

While some dominant actors, such as Amazon, Google, or Microsoft, and some public actors as intelligence agencies have the capacities to build large-scale AI systems entirely on their own, i.e., without purchasing external expertise, pre-trained models, data, or hardware resources, most actors do not. However, cloud computing platforms make those resources accessible and affordable to the many. Furthermore, as Keller et al. (2018, p. 8) note, "access to open source tools and frameworks for creating AI systems also play a part in the current wave of excitement. Tensorflow, Torch and Spark are examples of open source software libraries which […] have made the creation of AI systems – especially during research and development – significantly easier." Besides offering tools to develop and train AI systems, some providers of machine learning infrastructure also offer pre-trained machine learning models that can be incorporated into applications allowing to "score and classify new content right away" available (Microsoft, 2018). While some actors offer a multitude of services in the fields of hardware access, data preparation, model building, and production, there are also large numbers of specialized actors who only offer one or few services in one of these domains (Dhinakaran, 2020).

Thus, AI systems based on machine learning are often not monolithic applications developed by one actor or group of actors. Instead, they are complex socio-technical systems consisting of various technical components that are potentially developed,[4] managed, and operated by various independent actors or groups of actors. For instance, in the case of an AI system that aims to identify risk factors in patients' health records by detecting patterns learned from patient data, several tasks need to be considered: developing the underlying NLP capabilities, providing and preparing health records as training data, building a model that recognizes patterns based on this data, and using the system to classify or score unknown health records. All these tasks might be addressed by different actors. Moreover, due to business interests or strict data protection regulations related to health records, involved actors could be inclined not to share relevant information about the respective components of the system they

[4] Or, in case of data: generated or collected.

control or engage with. Regarding the data perspective, this challenge in healthcare is described well by Kemppainen et al. (2019).[5,6]

As Sect. 4 describes in detail, threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights can originate in various of these tasks. In some cases, they are even rooted in more than one of them. This raises the question of which of these actors are sensible addressees for obligations in AI regulation.

3. Two Frameworks for Assigning Obligations to Actors in AI Systems

Most recent European policy papers and regulatory proposals targeting AI systems recognize the system's considerable formative power. They view the introduction of AI systems as an opportunity to "stimulate new kinds of innovations that seek to foster ethical values" and to "improve individual flourishing and collective wellbeing" (HLEG-AI, 2019, p. 9). Yet, they also acknowledge the risks that AI systems pose for the realization of ethical values, the consideration of ethical principles, and fundamental rights (European Commission, 2020a). Here, recurring motives are risks for the respect for human decisions, self-determination, and agency; control over personal data; non-discrimination and fairness; accountability; technical robustness and safety; the rule of law; welfare systems; and democracy (see Datenethikkommission, 2019; European Commission, 2019, 2020a; HLEG-AI, 2019). Therefore, they propagate the creation of a regulatory framework that allows harnessing the potential of AI systems while simultaneously mitigating the risks associated with them. Developing such a framework requires addressing regulatory challenges that are familiar from other contexts as well as regulatory challenges that are specific to AI. For instance, a challenge common to the regulation of AI systems as well as the regulation of many other computer systems is the involvement of many actors in development processes, which makes it difficult to hold individual actors accountable (Nissenbaum, 1994). On the contrary, a challenge specific to AI systems

[5] The well-publicized case of "Watson Oncology," for example,exhibits many of these characteristics (Ross & Swetlitz, 2018; Strickland, 2019).

[6] Please note that the constellations of actors involved in AI systems are not uniform.Therefore, the features of the outlined case are not generalizable. It servesfor illustrative purposes only.

is the continuous learning "'in the wild,' that is, in uncontrolled real-world conditions" after deployment (Vallor & Bekey, 2017, p. 341).

The AI Act is a proposal for the regulation of artificial intelligence introduced by the EC seeking "to lay down harmonised rules for the development, placement on the market and use of AI systems" (Veale & Zuiderveen Borgesius, 2021, p. 2). The whitepaper *On Artificial Intelligence* is a document that sets out policy options on how to achieve "the twin objective of promoting the uptake of AI and of addressing the risks associated with certain uses of this new technology" (European Commission, 2020a, p. 1). Both the AI Act and the whitepaper propose a risk-based approach to regulating AI, i.e., to apply different governance measures depending on a risk level assigned to the application based on its application area, features, and purpose. While AI systems that are considered to pose an unacceptable risk are outright prohibited, especially in the case of high-risk applications and limited risk applications,[7] a large proportion of suggested measures take the form of obligations for regulated actors (European Commission, 2020a, 2021c; Veale & Zuiderveen Borgesius, 2021).[8] However, given the multitude of actors involved in the development, deployment, and operation of many AI systems, there are different approaches to assigning obligations to these actors.

The EC's whitepaper *On Artificial Intelligence* proposes a capability-based approach to assigning obligations. The whitepaper states that in the EC's view, "each obligation should be addressed to the actor(s) who is (are) best placed to address" the respective issue (European Commission, 2020a, p. 22).[9] However, the whitepaper does not outline how to determine which actor involved in an AI system is best placed to address an issue in detail (Borutta et al., 2020, p. 6). It just illustrates the approach by

[7] For further information on the risk-based approach and the classifications in the respective proposals, see European Commission (2021c, 3, 6, 13) and European Commission (2020a, p. 17).

[8] According to the AI Act, AI systems of two categories are considered high-risk. These are products "already covered by certain Union health and safety harmonisation legislation (such as toys, machinery, lifts, or medical devices)" Veale and Zuiderveen Borgesius (2021, p. 9), on the one hand, and AI systems for the use in further specified sensitive areas, on the other hand. Such sensitive areas are, for instance, biometric identification, law enforcement, and the administration of justice and democracy European Commission (2021a, 2021c).

[9] The whitepaper excludes questions of (civil) liability from this line of reasoning, arguing that it is not a premature judgement of question concerning "liability to end-users or other parties suffering harm and ensuring effective access to justice, which party should be liable for any damage caused" European Commission (2020a, p. 22).

suggesting that “[f]or example, while the developers of AI may be best placed to address risks arising from the development phase, their ability to control risks during the use phase may be more limited [in which case] the deployer should be subject to the relevant obligation” (European Commission, 2020a, p. 22). Further elaborations regarding the addressees of obligations only concern the geographic scope of the proposed regulation. They do not further specify by which criteria regulators should determine which actor is best placed to address a specific risk (Borutta et al., 2020; European Commission, 2020a, p. 22).[10] In the proposal for the AI Act, the EC departs from this view. It moves away from determining addressees of regulatory measures by evaluating their capability. Instead, it attempts to assign obligations to well-defined and clearly identifiable actors. Here, the focus is on “providers” and, to a lesser degree, “users” as the main addressees of obligations (Veale & Zuiderveen Borgesius, 2021). The EC defines “providers” as “a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge,”[11] and “users” as “any natural or legal person, public authority, agency or other body using an AI system under its authority, except where the AI system is used in the course of a personal non-professional activity” (European Commission, 2021c, pp. 39–40). With its approach to assign obligations to fixed addressees, the AI Act circumvents the necessity to engage with the possibly ambiguous setup of competencies and capabilities of actors involved in developing, deploying, and operating AI systems. However, Article 28 of the AI Act defines exceptions to this approach. According to Article 28, distributors, importers, users, and third parties are considered providers under the AI Act if “(a) they place on the market or put into service a high-risk AI system under their name or trademark; (b) they modify the intended purpose of a high-risk AI system already placed on the market or put into service; (c) they make a substantial modification to the high-risk AI system” (European Commission, 2021c, Art. 28).

[10] Such a criterion could be, for example, bearing the (least) cost for addressing a given risk (least (or cheapest) cost avoider approach, (cf. Calabresi, 2008)).

[11] The focus on providers can be explained in part by the fact that the AI Act draws heavily on existing European product safety regulation (Veale & Zuiderveen Borgesius, 2021).

4. On the Roots of Threats Posed by AI Systems

As stated in Sect. 1, AI systems can pose threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights. This section sets out several types of these threats that AI systems pose and traces their roots back to the various tasks in the process of developing, deploying, and operating AI systems and the actors involved in performing these tasks.[12] The list does not aspire to be exhaustive but merely demonstrates that these threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights can arise in the various tasks outlined in Sect. 2.

Purpose Setting: Some of the threats that AI systems pose are rooted in the setting of the system's purpose. The most salient issues discussed in the scholarly literature are use cases that are problematic from an ethical perspective, regardless of specific design decisions. First and foremost, these are AI systems for intentionally malign purposes such as "Prioritizing targets for cyber attacks using machine learning," "State use of automated surveillance platforms to suppress dissent" (Brundage et al., 2018), or attempting to mask intentional discrimination with ostensibly objective algorithms (Barocas & Selbst, 2016). Furthermore, while some use cases cannot be classified off-hand as based on malign intent, they nevertheless are posited at least at the edge of moral dubiousness because the prospect of deploying such a system raises severe ethical concerns regardless of specific design decisions. Prime examples of such systems are, for instance, lethal autonomous weapons systems (Horowitz, 2016).[13]

The AI Act addresses some of these purposes already by assigning them to an "unacceptable risk" category. Within this category, the AI Act does not assign obligations to specific actors involved in the development, deployment, and use of the system but operates with "outright or qualified prohibitions" for respective applications. The current proposal for the AI Act "contains four prohibited categories, three prohibited in their entirety (two on manipulation, one on social scoring); and the last, 'real-time' and 'remote' biometric identification systems prohibited except for

[12] While the list is loosely oriented to the sequence of a development process, this is not meant to suggest that the development processes of most AI applications are linear. Often, prototypes are taken into an early deployment and subsequently evaluated and further refined during their use "in the wild."

[13] For a vivid illustration of the issue of morally problematic use cases and the importance of ethically sound purpose setting, see Keyes et al. (2019).

specific law" (Veale & Zuiderveen Borgesius, 2021, p. 3). In these cases, the framework discussed in this article does not apply.

Data Management and Data Preparation: Many of the most controversially discussed threats posed by AI systems based on machine learning have their roots in data management and data preparation. This is because the data used for training an AI system severely impacts how it determines its decisions later on. As Barocas and Selbst (2016) explain, "the data that function as examples [...] train the model to behave in a certain way." Data preparation in this context "involves preparing, labeling, and cleaning the data to be used for models" (Dhinakaran, 2020). The main issues that can arise here can be divided into four categories: (1) the use of inaccurate data, (2) the use of nonrepresentative and insufficient data, (3) data containing pre-existing societal biases, and (4) the use of unsecured, protected data.

In many cases in which inaccurate data leads to AI systems posing threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights, inaccurate labeling is at the core of the problem. According to Barocas and Selbst (2016), "labeling examples is the process by which the training data is manually assigned class labels" and further that "the labels applied to the training data must serve as ground truth" for the system. Thus, inaccurate labels lead to a skewed ground truth. For instance, in the field of AI systems for medical image classification, "[i]mage labels are annotations performed by medical experts such as radiologists" (Willemink et al., 2020). Inaccurate training data (or, more specifically, inaccurate labels in the training data) here come about if medical experts categorize and annotate images incorrectly. Inaccurate data can also be deliberately injected into training data to manipulate an AI system's decisions. For instance, in recommender systems, inaccurate recommendations (e.g., fake product recommendations) made by users can shift the recommendations that a system provides. As Milano et al. (2020) note, "providing inaccurate or irrelevant recommendations directly harms a user by reducing the utility that they derive from the recommended option." Other forms of harm can occur if attackers manipulate the training data of AI systems that are applied in other domains.

Besides inaccurate data, nonrepresentative data and data containing pre-existing societal biases (Friedman & Nissenbaum, 1996) play a crucial role in the threats that AI systems pose, and that can be attributed to the preparation and management of data. Many of the threats that are discussed under the heading "algorithmic bias" fall in this category. Nonrepresentative data here refers to data sets that omit to sufficiently take certain groups into account. For instance, image recognition software, in many cases, is trained with pictures of predominantly light-skinned persons. In an illustrative case, the error rates for identifying individuals of an image recognition tool developed by Amazon differed extensively between population groups, particularly lighter- and darker-skinned individuals. It mislabeled especially darker-skinned women disproportionately often (Arbel, 2019). From an ethical perspective, the problem is exacerbated by that "[e]rrors of this sort may befall historically disadvantaged groups at higher rates because they are less involved in the formal economy and its data-generating activities, have unequal access to and relatively less fluency in the technology necessary to engage online, or are less profitable customers or important constituents and therefore less interesting as targets of observation" (Barocas & Selbst, 2016). Furthermore, they conclude that this does not only affect the "quality of individual records of members of these groups be poorer as a consequence, but these groups as a whole will also be less well represented in datasets, skewing conclusions that may be drawn from an analysis of the data."

The same issues can arise even if the data is representative, i.e., reflects the overall population, but is insufficient in its extent regarding a certain group. For instance, from a technical perspective, the error of image recognition software would deliver equally problematic results if the training data was representative of the overall population, but the sample of persons with a given skin color would be minuscule in the overall population. It can be expected that the error rate in recognition of images of persons with that skin color would be higher than in other groups, not because it is underrepresented in the training data in that it "deviates from the actual population statistics" (Danks & London, 2017), but that that there is insufficient training data for a certain subset of the population to achieve high-quality outcomes, leading to differential treatment of different population groups.

Moreover, even if the training data is accurate, the data sample is representative of the overall population, and there are sufficient datasets concerning all relevant subgroups, it can still be skewed regarding a moral standard, leading to biased decisions if used as training data for AI systems. Such issues arise "if particular groups ([e.g.,] based on race, religion, ethnicity etc.) have historically suffered disadvantage" (Yeung, 2019,

p. 32), and this fact is reflected in the data. For instance, several cases show that AI systems used in hiring processes often disfavored women and racial minorities even if the applicants from these groups had “credentials otherwise equal to other applicants” because the training data was based on historical hiring practices and mirrored existing discrimination (Barocas & Selbst, 2016; see also Dastin, 2018; Lowry & Macpherson, 1988; Yeung, 2019). In these cases, AI systems replicated and reinforced historical biases and perpetuated “injustice against disadvantaged groups and associated stereotypes and stigmatization” (Yeung, 2019, pp. 32–33), even though the data was neither incorrect nor misrepresenting the status quo. Instead, the underlying problem in these cases is that the “relevant moral standard”—the equal or fair treatment of women and racial minorities—“is different from the current empirical facts” (Danks & London, 2017, p. 4693).

Additionally, making decisions on individual human beings requires working with some form of personal data. Here, especially personally identifiable information can lead to privacy issues. Moreover, the prioritization, classification, association, or filtering of individuals by AI systems can create sensitive insights, even if originally less sensitive data is used as an input. In such cases, “[p]ersonal harms emerge from the inappropriate inclusion and predictive analysis of an individual’s personal data” (Crawford & Schultz, 2014, p. 94). In a prominently discussed case, the retail chain Target used customer data to make predictions on whether customers were pregnant. It then forwarded this information to marketers to target the respective customers with relevant products, even in cases where the customers did not announce their pregnancy publicly yet (Duhigg, 2012; Hill, 2012). In essence, Target’s actions “resulted in the unauthorized disclosure of personal information” (Crawford & Schultz, 2014). Thus, AI systems bring about novel privacy-related issues, such as the “predictive privacy harms” (Crawford & Schultz, 2014) from the Target case.

Model Development: Also, activities during the development of an AI system’s decision model can cause threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights. While many process steps are associated with model development (see Dhinakaran, 2020), most of the threats discussed in the pertinent literature concern the setting of target variables or feature selection.

The setting of target variables is partly determined by the purpose of the AI Application. Hence, the threats posed by AI systems outlined under “Purpose Setting” can manifest here if problematic purposes are operationalized as target variables in model development. However, target variables are not only determined by the purpose of an application. Target variables also contain further requirements to the output of a system that are of ethical relevance. Among these are, for instance, fairness metrics that determine how and to what degree fairness is considered in the differential treatment of various groups. As Binns (2018) points out, there are various metrics for, e.g., fairness, "including; 'accuracy equity', which considers the overall accuracy of a predictive model for each group; 'conditional accuracy equity', which considers the accuracy of a predictive model for each group, conditional on their predicted class; 'equality of opportunity', which considers whether each group is equally likely to be predicted a desirable outcome given the actual base rates for that group; and 'disparate mistreatment', a corollary which considers differences in false positive rates between groups." To complicate matters, the different measures are often "mathematically impossible to satisfy simultaneously except in rare and contrived circumstances, and therefore hard choices between fairness metrics must be made before the technical work of detecting and mitigating unfairness can proceed" (Binns, 2018; see also Kleinberg et al., 2016). Failure to recognize and act on such concerns during the model development thus can cause severe issues concerning fairness, other ethical values, principles, or fundamental rights.

The process of "feature selection" refers to the process of making choices about what attributes in data sets to observe and subsequently fold into an analysis (Barocas & Selbst, 2016). It can pose threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights if features concern morally or legally sensitive attributes and make them a determining factor in a system's decision-making model. This can either manifest in the selection of a feature that directly represents sensitive attributes as race or gender or via proxy information that serve as a placeholder for such attributes (Danks & London, 2017, p. 4696). As Crawford and Schultz (2014, p. 100) note, by using proxies, AI systems can "circumvent anti-discrimination enforcement mechanisms by isolating correlative attributes that they can use as a proxy" for protected attributes.

Moreover, if data that causally relates to relevant variables is challenging to obtain, models can take correlating data into account as a proxy for unavailable data. This creates the risk of sensitive data being used where it "might be capable of serving as an informational proxy for a morally unproblematic, though hard to measure, variable or feature" (Danks & London, 2017, p. 4696). In some cases, these variables "have a very high degree of predictive value ([i.e.,] statistical relevance)" (Yeung, 2019, p. 27). For this reason, using sensitive attributes as features can be appealing to users even if they are morally problematic. However, individuals that AI systems make decisions on still have a "legitimate interest in not being evaluated and assessed based on considerations that are not causally relevant to the decision" (Yeung, 2019). As a result, the interests of operators and those affected by AI systems may differ extensively, leading to conflicts between different stakeholder groups.

Furthermore, as Barocas and Selbst (2016) point out, many cases discussed in pertinent literature suggest that model developers often settle for proxies which serve as a "highly imperfect basis upon which to predict" other features of an individual that are causally relevant for a decision. Prominently discussed cases are, for instance, the use of skin color as a proxy for the likelihood of an individual having a criminal record (Barocas & Selbst, 2016; Strahilevitz, 2008) and gender as a proxy for traits that correlate with job performance (Danks & London, 2017).

Other threats posed by feature selection arise because, in many contexts, not all groups are equally represented in the set of selected features. Barocas and Selbst (2016) elaborate: "Members of protected classes may find that they are subject to systematically less accurate classifications or predictions because the details necessary to achieve equally accurate determinations reside at a level of granularity and coverage that the selected features fail to achieve."

While the sources of the discussed issues are already rooted in the training data, feature selection can fail to take these issues of the training data into account. If the selection of problematic features is unavoidable, the model can integrate mechanisms that counter or offset adverse outcomes of an AI system's decisions. Here, fairness metrics can also play a role here.

Deployment, Use, and Refinement: Lastly, threats that AI systems pose can also be rooted in the way they are deployed, used, and refined. Regarding AI systems' deployment, especially its embedding in its socio-technical environment can be a cause of concern. In the technical domain, poor technology-environment design can bring about malfunctions, leading to erroneous decision-making with potentially severe consequences for affected individuals. Similarly, the deployment of an AI system into a social environment with characteristics differing from the ones assumed during its development can cause threats to individuals or society. For instance, Friedman and Nissenbaum (1996) observe that a mismatch between users and system design can occur if “the population using the system differs on some significant dimension such as expertise or cultural background from the population assumed as users in the design” and lead to biased system behavior. This holds true also for AI systems. Recalling the threats that can arise in data preparation and data management illustrates why: while a training data set might not be biased relative to a standard of statistical distribution in one context, it might be in another (Danks & London, 2017, pp. 4692–4693). Therefore, the threats described above, especially the ones stemming from regard to data preparation and data management, can unexpectedly occur if a system is deployed in a context that it was not designed for. Control mechanisms put in place for the intended context of use may fail in a different context.

Moreover, the continuous (re-)development, refinement, and feedback loops between users, developers, and the system give room to novel types of threats. A continuous expansion of training data and, building on that, a constantly evolving decision-making model require a continuous evaluation of the ethical soundness of AI systems over time. Even if an AI system is considered unobjectionable or harmless at one point in time, an evolvement of the system can lead to model instability and performance degradation (Cheatham et al., 2019).

Especially in security-related contexts, “asymmetric feedback” can be a source of performance degradation in systems that integrate continuous learning (O'Neil, 2016). Asymmetric feedback emerges if the setting that a system is placed in only allows for unilateral feedback. For instance, as Zweig et al. (2018, p. 193) note, “a criminal offender who is not released on bail on the recommendation of an ADM system has no way to prove that he would not have recidivated.” In the case of a binary decision like this, the system thus only can get feedback on one type of decision and only learn from one type of mistake, leading to over-specialization in one direction while not recognizing and reacting to mistakes in the other direction (Zweig et al., 2018).

Another way cause for performance degradation of AI systems is learning from interactions with human actors that—intentionally or unintentionally—feed the system problematic input. An often-cited case illustrating this issue is the chatbot

"Tay," released by Microsoft in 2016 to be shut down after only one day, because the model "quickly turned offensive and abusive after interacting with Twitter users" (Neff & Nagy, 2016, p. 4921). While the abusive behavior of the bot did not directly stem from the developer's actions—it "echoed the racism and harassment that was fed into it" (Neff & Nagy, 2016) by social media users—the developers were accused of being accountable for not considering the possibility of such a performance degradation and putting "in place additional safeguards and testing procedures" (Wolf et al., 2017).

However, while the intervention of human actors can be necessary to deal with erroneous or biased AI systems, and overriding or intervening in decisions can be necessary, it can also be a source of further bias. As Cheatham et al. (2019) note, "human judgment can also prove faulty in overriding system results," leading to a biased or otherwise unethical decision and, potentially, feeding these decisions as new input data into the system. This, in turn, can result in future replication, i.e., similar—and similarly problematic—decisions made by the AI system.

Conversely, "automation bias" (cf. Skitka et al., 2000)—a tendency to trust or rely on technical artifacts to a higher degree than is warranted—can also be problematic as it leads to operators of automated systems paying "insufficient attention to monitoring the process and to verifying the outputs of the system" (Simon et al., 2020, pp. 12–13). This issue can be exacerbated if operators receive insufficient training to adequately assess an AI system's output as well as its reliability and do not "recognize when systems should be overruled" (Cheatham et al., 2019).

5. Discussion

Based on the analyses in the previous sections, this section discusses the merit of the shift from the capability-based framework outlined in the EC's whitepaper *On Artificial Intelligence* to the framework based on fixed addressees outlined in the AI Act. The evaluation of the frameworks rests on determining to what extent they are able to deal with three challenges that are derived from the findings in previous sections.

Ambiguity Regarding Which Actors Are Best Placed to Address Risks or Negative Consequences: It is inherently ambiguous which actor is best placed to address threats posed by AI systems. The EC's proposal that "each obligation should be addressed to the actor(s) who is (are) best placed to address any potential risks" (European Commission, 2020a) therefore cannot be translated into practice straightforwardly. This is the case even if not only the risk or negative consequence itself but also its root of a threat posed by an AI system is in plain view. This is because identifying the root of a threat does not directly provide any information on who is best placed to address it. If, for instance, an AI system has proven to be biased and problematic features of the training data have been identified as the source of the issue, it can be resolved within the remit of different actors.

Firstly, the actors engaged in data preparation and data management can modify, replace, or delete data points in the training data that—in their aggregate—have shown to be biased. Secondly, the actors engaged in model development "can use a bias in the algorithmic processing to offset or correct for the data bias, thereby yielding an overall unbiased system" (Danks & London, 2017, p. 4695). Thirdly, the deployment and use of an AI system can be adjusted by "restrict[ing] the scope of operation for the system in question so that there is no longer a mismatch in system performance and task demands" or, in case of decision support systems, by the user "deliberately employ[ing] a compensatory bias" instead of taking action "solely on the basis of the algorithm output" (Danks & London, 2017, p. 4695). Thus, in a capability-based framework, this ambiguity creates the need for regulatory bodies to further specify which actors should be addressed. Approaches such as the cheapest cost avoider principle could reduce this ambiguity by providing clear criteria. However, such a principle can only be applied to individual cases and does not provide generalizable rules for assigning obligations.

On the contrary, the AI Act's framework appears to be appropriate to address the challenge of assigning obligations to actors without regulatory gaps arising due to ambiguous addressees of obligations. By disentangling obligations from capability or judgments on who is "best placed," the proposal for the AI Act allows assigning obligations without engaging with the actor constellations in individual AI systems and the respective actor's capabilities. Yet, while the AI Act's approach results in clearer attribution of obligations, these are not strictly linked to the actual causes and solutions of the respective problems. Therefore, the actors addressed by the AI Act need to establish this link by identifying actors within the respective system that are capable of, e.g., providing documentation or securing ethically relevant technical features of the system and ensuring that these actors support them in fulfilling their obligations. Therefore, the question of which actor is capable of or best placed to address a threat posed by an AI system is not irrelevant in the framework underlying the AI Act. Addressing this issue is merely delegated from the regulatory authority to providers of AI systems.

However, the AI Act recognizes that providers cannot fulfill their obligations under some circumstances. As noted in Sect. 3 of this article, Article 28 determines that if distributors, importers, users, or third parties modify the intended purpose of a high-risk AI system or make substantial modifications to it, they take the role of the provider of that AI system henceforth. All obligations of the original provider are transferred to them (European Commission, 2021c, Art. 28). Here, the AI Act does not strictly follow the framework of fixed addressees but engages in redefining roles and reassigning obligations based on specific actions by involved actors. This deviation from assigning obligations based on a framework based on fixed addressees poses similar challenges as the challenges for the capability-based framework mentioned above: ambiguities arise, which are hard to address due to the complexity and lack of uniformity of AI systems. Smuha et al. (2021, p. 28) raise the question if there are "cases in which a user may legitimately 'misuse' a particular AI system to protect fundamental rights (could the user then change the intended purpose of an AI system without incurring the obligations of a provider under Article 28) [and, if so] who decides these thresholds?" Furthermore, as Ebers et al. (2021, p. 597) note, in the case of AI systems that are "used for many different purposes (general-use AI systems), there may be circumstances where such an AI technology gets integrated into a high-risk system, without the provider having any or only limited influence over the compliance obligations of high-risk AI systems." Here, the question arises if such an integration is considered misuse or in accordance with the purpose of the system. [14] These ambiguities can be illustrated revisiting the example of applying a multi-purpose NLP system in a healthcare setting introduced in Sect. 2. Which obligations the AI Act assigns to the provider of the system and which it assigns to the user of the system depends on various factors. First, while this circumstance has been criticized (see, e.g.,

[14] In its reply to the AI Act, Google makes a similar argument claiming that the rules for shifting obligations from providers to other involved actors lack clarity and that "companies will be forced to take a conservative position, imposing a significant chilling effect on the release of general-use APIs and OSS until the issue is resolved in the courts" (Google, 2021, p. 21).

Ebers et al., 2021, p. 594), applying AI systems in sensitive contexts such as healthcare does not automatically qualify this system as a high-risk system. Therefore, the obligations to providers, users, and other actors defined in chapter 2 of the AI Act do not apply by default. However, if the purpose of the system meets the criteria defined in Annex II and Annex III of the AI Act (see Sect. 3), the AI Act assigns additional obligations for high-risk AI systems to the respective actors. For instance, if the AI system is a medical device, it would meet the criteria defined in Annex II. If this is the case, by default, the majority of obligations falls on the provider, and only a few requirements are assigned to the user (e.g., ensuring "that input data is relevant in view of the intended purpose of the high-risk AI system" and monitoring "the operation of the high-risk AI system on the basis of the instructions of use") (European Commission, 2021c, Art. 29). However, two further factors determine this attribution of obligations. On the one hand, in the case of healthcare data, providers will often not have access to the relevant data (e.g., patient records) to fulfill their obligations (Kemppainen et al., 2019). Based on the distribution of control over the data between user and provider, the obligation to "ensure that input data is relevant in view of the intended purpose of the high-risk AI system" is assigned to one or the other (European Commission, 2021c, Art. 29). On the other hand, the question of whether integrating a multi-purpose AI system in a high-risk application is to be considered a modification of the purpose of the system or a misuse of the system remains relevant. As mentioned above, if either is the case, the entirety of obligations originally assigned to the provider would be instead assigned to the user (European Commission, 2021c, Art. 28).

The Insufficient Informational Basis for Addressing Threats Posed by AI Systems: To address threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights posed by AI systems, involved actors need an informational basis to do so. For instance, if model developers are supposed to decide on whether to integrate a compensatory bias in an AI system's decision model to account for an initial bias in the training data, they need a solid informational basis provided by the actors responsible for data collection, preparation, and management on relevant features of the respective datasets. Thus, the feasibility of meeting obligations is often dependent on receiving information from actors who are better placed to assess features of technical components, use cases, and application contexts (Digital Europe, 2021, p. 5). However, the more independent actors are involved in developing, deploying, and operating an AI system, the less likely it is that the necessary information transfer will occur.

Moreover, recent trends in the development of AI systems stand in the way of the necessary disclosure and consideration of information flowing in both directions: from the collection and management of training data, over model development, to deployment and use as well as the other way around. This flow of information is often limited due to the involved actors' business interests. While algorithms and data are non-rivalrous goods, openly sharing them can still lead to a "loss of advantage over competitors" (Keller et al., 2018, p. 11). If little information is shared about the training data or algorithms used to train an AI system's decision model, it becomes increasingly difficult for actors who integrate the model into end-user applications or actors who deploy and operate them to react adequately to ethically problematic properties of the system.

Conversely, actors who are involved in managing training data or developing decision models in many cases lack or cannot fully account for information on societal features of the context of use of an end-user application in order to adjust the AI system to this context and, e.g., avoid bias (see Friedman & Nissenbaum, 1996). This is because the context of use is often deliberately not fully defined to allow components of a system to be used in more than one application context. Pre-trained models are a prime example of this. Empirical research has already linked existing pre-trained models to bias (Webster et al., 2020) as well as security-related issues (Gu et al., 2019). It needs to be emphasized, however, that the problem is not limited to the use of pre-trained models but is more general.

While the proposal for the AI Act requires providers of AI systems of "high-risk" category to "establish a sound quality management system, ensure the accomplishment of the required conformity assessment procedure, draw up the relevant documentation and establish a robust post-market monitoring system" (European Commission, 2021c, p. 31), it refrains from defining the scope of information obligations among the other actors involved in the system. The information and documentation obligations are directed primarily at providers. Nevertheless, establishing a structure for sharing information among the involved stakeholders is, in practice, a prerequisite for providers to fulfill the AI Act's obligations. For instance, if providers of the system are not themselves directly in charge of data management, data preparation, or model development, they can hardly establish a quality management system (European Commission, 2021c, Art. 17), provide technical documentation (European Commission, 2021c, Art. 11), or ensure adequate data governance (European Commission, 2021c, Art. 10) without entering into an intensive exchange of information with the actors in charge of the respective tasks. This is because, as Digital Europe (2021, p. 5) notes, questions such as "what is relevant and representative at a given time when developing the AI system will vary based on the use case," and in many cases, providers need to rely on users to assess use cases. Thus, to ensure that it is feasible to fulfill their obligations, providers need to ensure that development practices (e.g., the use of pre-trained models) or business practices (e.g., restricting access to resources such as data or algorithms) do not obstruct necessary information sharing between the involved actors.

The capability-based framework—as presented in the whitepaper *On Artificial Intelligence*—does not define any requirements for information sharing among involved actors or between the involved actors and regulatory authorities. Nevertheless, the approach is affected by insufficient data-sharing practices because whether or not a given actor possesses access to information determines if it is well placed to address a threat posed by an AI system. For instance, actors in model development could be well posited to address a given bias that results from skewed training data by integrating a compensatory bias. However, one could only meaningfully describe these actors as well-positioned to address the respective issue if information about the relevant features of the training data is shared with them. Actors controlling data could refrain from sharing such information, for instance, based on business interests (Keller et al., 2018) or, as described in the healthcare case, due to data protection regulation. Therefore, a capability-based approach applied in practice needs to spell out either an information-sharing ruleset or it needs to evaluate actual information-sharing practices on a case-by-case basis to determine which actor is best placed to address an issue. Both approaches would involve a major regulatory burden if applied by regulatory authorities to prevent regulatory gaps due to ambiguous addressees of obligations. The elaborations in the whitepaper *On Artificial Intelligence* do not engage with this issue and therefore leave this central issue unaddressed.

Thus, the framework introduced in the proposal for the AI Act based on fixed addressees avoids further central problems that occur in the capability-based framework proposed in the whitepaper *On Artificial Intelligence*. While it does not provide a clear path to how information sharing should be structured among the involved actors, it establishes a well-defined, clearly identifiable actor—the provider—who is the main addressee of most obligations. Thereby, the proposal for the AI Act makes it possible to delegate the micromanagement of information sharing without allowing regulatory gaps due to ambiguous addressees of obligations to arise.

Systemic, Cumulative Effects of AI Systems: Finally, some threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights posed by AI systems are not rooted in one specific application, let alone individual actions during the process of developing, deploying, and operating it. Most importantly, these are issues concerning the "cumulative effect from widespread and systematic reliance on algorithmic decision-making [which] could erode and destabilize the core constitutional, moral, political, and social fabric upon which liberal democratic societies rest and upon which our shared values are rooted" (Yeung, 2019, pp. 41–42). For instance, the use of AI systems by social networks has been criticized for increasing political polarization (see, e.g., Hao, 2021), whereas the use of AI systems by health insurances is being critically examined regarding whether individually justified differentiations can lead to a loss of solidarity in society (see, e.g., Datenethikkommission, 2019). Furthermore, the capacity of AI systems to make inferences about individual's intimate aspects of life and decisions that determine their future based on data that individuals produce by going on with their everyday life "may have a corrosive chilling effect on our capacity to exercise our human rights and fundamental freedoms" (Yeung, 2019, p. 36).[15,16]

The question of who bears responsibility for the systemic, cumulative effects of the widespread use of AI systems for society is not addressed by either framework. In the case of the AI Act, this might be caused by that it is modeled after EU product law, especially regulation concerning product safety (Veale & Zuiderveen Borgesius, 2021,

[15] For a more exhaustive discussion, see Yeung (2019).

[16] Systemic, cumulative effects of AI systems are not listed in Sect. 4, as Sect. 4 attempts to trace the roots of threats back to the various tasks in the process of developing, deploying, and operating an individual AI system.

p. 3). Thus, irrespective of which framework is applied, further policy considerations are required to address threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights that fall outside the scope of the respective framework.

6. Conclusion

Currently, a broad range of academic literature and numerous (European) policy papers, as well as regulatory proposals, discuss how AI systems can and should be regulated. Within this discourse, one key challenge that is discussed is how to determine which of the actors involved in developing, deploying, and operating AI systems should be obliged to address threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights such systems can pose. The present article contributes to this discourse by discussing the appropriateness of the frameworks to assign such obligations to involved actors proposed by the EC in the whitepaper *On Artificial Intelligence* and the proposal for the AI Act, respectively.

To do so, this article first provides an overview of the different tasks that exist in the process of developing, deploying, and operating AI systems and the actors involved in performing these tasks (Sect. 2). Then, it introduces the frameworks to assign obligations outlined in the EC's whitepaper *On Artificial Intelligence* and the AI Act, respectively (Sect. 3). Subsequently, the article links the threats posed by AI systems to the various tasks in the process of developing, deploying, and operating AI systems (Sect. 4). Finally, it discusses challenges for the two frameworks and the merit of the shift from the capability-based framework outlined in the EC's whitepaper *On Artificial Intelligence* to the framework based on fixed addressees outlined in the AI Act (Sect. 5).

The capability-based framework—targeting actors who are "best placed" to address threats—suffers from the fact that one threat, for instance, bias against a protected group, can have various roots and paths to resolve. Therefore, which involved actor is most capable of or best placed to address a threat posed by an AI system remains highly subjective. Furthermore, threats posed by AI systems often do not emerge as a result of one actor's activity but due to an insufficient flow of information among several involved actors. An actor who, in theory, is capable of addressing a threat posed by an AI system does, in practice, often not have access to information that allows it to recognize this circumstance and react adequately to it. Both concerns can lead to a diffusion of responsibility as they give involved actors leeway to reject obligations to address a specific threat assigned to them. Therefore, if the capability-based framework for regulating AI systems is applied, it would require extensive micromanagement of regulatory authorities in assigning obligations. As there are no more tangible proposals dealing with the challenges identified in Sect. 5, the capability-based framework, as outlined in the whitepaper *On Artificial Intelligence*, does not provide an appropriate path to assigning obligations to actors involved in developing, deploying, and operating AI systems. While providing additional criteria for what “best placed” means in practice could reduce ambiguity, it would not reduce the need for extensive micromanagement by regulatory authorities.

The framework based on fixed addressees outlined in the proposal for the AI Act is less affected by both challenges: the ambiguity regarding which actor is best placed to address a given threat posed by an AI system, and the insufficient informational basis for actors to address such threats. By obliging actors who place a product on the market to ensure that it meets a given set of criteria, this framework avoids the necessity for regulators to engage in-depth with the actor constellations within a given AI system, as the obligations are simply assigned to a well-defined and clearly identifiable actor. The responsibility to gather the necessary information and ensure certain system properties, even in ambiguous setups, is delegated to these actors. By relying on this framework, the proposal for the AI Act resolves some of the core problems of earlier policy papers and regulatory proposals and is thus more appropriate in this crucial respect.

However, in cases in which providers are not in charge of the whole process of developing, deploying, and operating AI systems, they might need to rely on additional actors involved in an AI system to cooperate to fulfill obligations such as establishing a quality management system (European Commission, 2021c, Art. 17), providing technical documentation (European Commission, 2021c, Art. 11), or ensuring adequate data governance (European Commission, 2021c, Art. 10). To do so, providers need to identify actors capable of providing information and carrying out modification and oblige them to do so. In practice, this could result in that it is unfeasible for providers to cooperate with actors who rely on engineering and business practices that are common in the development, deployment, and operation of AI systems but are incompatible with such requirements. Affected could be, for instance, the use of pre-trained models or business practices built around disclosing little information about the training data.

Yet, since the proposal for the AI Act categorizes the AI systems in question as posing a high risk or bringing about negative consequences for individuals or the society, the higher weighting of ensuring that no regulatory gaps arise over specific engineering and business practices is only consequential. Accordingly, the EC's shift from a framework focused on capability to a framework focused on fixed addressees is appropriate in that it ensures that there is a well-defined and identifiable actor that obligations can be assigned to.

Nonetheless, just like capability-based approaches for assigning obligations, the AI Act's framework based on fixed addressees aims at addressing threats posed by individual AI systems. It does not consider cumulative effects of AI systems resulting from a "widespread and systematic reliance on algorithmic decision-making" (Yeung, 2019). Thus, irrespective of the selected framework, this issue must be addressed through alternative policy considerations. Moreover, it is important to note that how a regulatory proposal deals with the challenge of determining sensible addressees for the respective obligations is by no means the only factor that determines its appropriateness. Both the proposal for the AI Act as well as the whitepaper *On Artificial Intelligence* have been criticized for other reasons, as, e.g., the appropriateness of the risk categorization, the appropriateness of risk-based approach in general, a wide scope for interpretation, an extensive bureaucratic burden, relying heavily on (self-) conformity assessments and proportionality assessments, and (further) regulatory blind spots (see, e.g., Borutta et al., 2020; Hoffmann, 2021; Smuha et al., 2021; Veale & Zuiderveen Borgesius, 2021). Therefore, this article should be understood as a contribution to the discourse on the respective proposal's appropriateness regarding the framework for assigning obligations to actors involved in developing, deploying, and operating AI systems and not as an exhaustive assessment of the respective proposals.

Future research and regulation can address some of the outlined issues for assigning obligations in AI regulation. Regarding the capability-based framework, developing criteria for what "best placed" means in practice is crucial. This could help to explore whether a more elaborate form of the framework would be a feasible alternative approach to assigning obligations. The AI Act, however, has to be complemented by further regulation. The EC is already planning to address (civil) liability issues “related to new technologies, including AI systems” that it did not address in the AI Act such as “revisions of the sectoral safety legislation and changes to the liability rules” (European Commission, 2021b, p. 1). Future research should, therefore, investigate to what extent the frameworks and concepts discussed in this article can be transferred to this context. Furthermore, the EC needs to address cumulative effects of AI systems on society with additional regulation.

Acknowledgements

Open Access funding enabled and organized by Projekt DEAL. Mattis Jacobs and Judith Simon acknowledge financial support from the German Federal Ministry of Education and Research in the project GOAL–About algorithmic behaviour control and artificial intelligence under the reference 01IS19020.

Publication’s References

Arbel, T. (2019). *Researchers say Amazon face-detection technology shows bias*. https://abcnews.go.com/Technology/wireStory/researchers-amazon-face-detection-technology-shows-bias-60630589?cid=social_twitter_abcn

Barocas, S., & Selbst, A. D. (2016). Big data’s disparate impact. *Calif. L. Rev.*, *104*, 671.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, *81*, 1–11.

Borutta, Y., Haag, M., Hoffmann, H., Kevekordes, J., & Vogt, V. (2020). *Fundamentalkritik des White Papers und des Datenstrategiepapiers der EU-Kommission vom 19. Februar 2020*. https://goal-projekt.de/wp-content/uploads/2020/03/Fundamentalkritik-1.pdf

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., . . . Amodei, D. (2018). *The Malicious Use of Artificial*

Intelligence: Forecasting, Prevention, and Mitigation. https://arxiv.org/pdf/1802.07228

Calabresi, G. (2008). The Cost of Accidents: A Legal and Economic Analysis. Yale University Press.

Cheatham, B., Javanmardian, K., & Samandari, H. (2019). *Confronting the risks of artificial intelligence*. McKinsey Quarterly. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence

Crawford, K., & Schultz, J. (2014). Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review*, *55*(1), 93–128. https://heinonline.org/HOL/P?h=hein.journals/bclr55&i=93

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. In F. Bacchus & C. Sierra (Eds.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4691–4697). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2017/654

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

Datenethikkommission. (2019). *Gutachten der Datenethikkommission*. https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf

Dhinakaran, A. (2020). *The AI Ecosystem is a MESS: Why is it impossible to understand what AI companies really do?* Towards Data Science. https://towardsdatascience.com/the-ai-ecosystem-is-a-mess-c46bdfbf43e4

Digital Europe. (2021). *DIGITALEUROPE's initial findings on the proposed AI Act*. Digital Europe. https://www.digitaleurope.org/wp/wp-content/uploads/2021/08/DIGITALEUROPEs-initial-findings-on-the-proposed-AI-Act.pdf

Duhigg, C. (2012). *How Companies Learn Your Secrets*. The New York Times. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H., & Steinrötter, B. (2021). The European Commission's Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). *J*, *4*(4), 589–603. https://doi.org/10.3390/j4040043

European Commission. (2019a). *Building Trust in Human-Centric Artificial Intelligence*. https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

European Commission. (2020a). *On Artificial Intelligence - A European approach to excellence and trust: Whitepaper*. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission. (2021a). *ANNEXES to the Proposal for a Regulation of the European Parliament and of the Council LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF

European Commission. (2021b). *COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT Accompanying the Proposal for a Regulation of the European Parliament and of the Council LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. https://eur-lex.europa.eu/resource.html?uri=cellar:0694be88-a373-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

European Commission. (2021c). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Google. (2021). *Consultation on the EU AI Act Proposal: Google's submission*. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2662492_en

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, *7*, 47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068

Hao, K. (2021). *The Facebook whistleblower says its algorithms are dangerous. Here's why*. MIT Technology Review. https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/

Hill, K. (2012). *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did*. Forbes. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=20f35caa6668

HLEG-AI. (2019). *Ethics guidelines for trustworthy AI*.

Hoffmann, H. (2021). Regulierung der Künstlichen Intelligenz: Fundamentalkritik am Verordnungsentwurf zur Regulierung der Künstlichen Intelligenz der EU-Kommission vom 21. 4. 2021. *Kommunikation & Recht*, 369–374.

Horowitz, M. C. (2016). The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons. *Daedalus*, *145*(4), 25–36. https://doi.org/10.1162/DAED_a_00409

Keller, J. R., Chauvet, L., Fawcett, J., & Thereaux, O. (2018). *The role of data in AI business models*. Open Data Institute. https://theodi.org/wp-content/uploads/2018/04/376886336-The-role-of-data-in-AI-business-models.pdf

Kemppainen, L., Pikkarainen, M., Hurmelinna-Laukkanen, P., & Reponen, J. (2019). Data Access in Connected Health Innovation: Managerial Orchestration Challenges and Solutions. *Technology Innovation Management Review*, *9*(12), 43–55. https://doi.org/10.22215/timreview/1291

Keyes, O., Hutson, J., & Durbin, M. (2019). A Mulching Proposal. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). ACM. https://doi.org/10.1145/3290607.3310433

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores*. https://arxiv.org/pdf/1609.05807

Krafft, T. D., Zweig, K. A., & König, P. D. (2020). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance*, *104*, 119–136. https://doi.org/10.1111/rego.12369

Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal (Clinical Research Ed.)*, *296*(6623), 657–658. https://doi.org/10.1136/bmj.296.6623.657

Microsoft. (2018). *Pre-trained machine learning models for sentiment analysis and image detection*. Microsoft. https://docs.microsoft.com/en-us/machine-learning-server/install/microsoftml-install-pretrained-models

Milano, S., Taddeo, M., & Floridi, L. (2020). Ethical aspects of multi-stakeholder recommendation systems. *The Information Society*, 1–11. https://doi.org/10.1080/01972243.2020.1832636

Neff, G., & Nagy, P. (2016). Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication*.

Nissenbaum, H. (1994). Computing and accountability. *Communications of the ACM*, *37*(1), 72–80. https://doi.org/10.1145/175222.175228

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First paperback edition). Broadway Books.

Quan, X. I., & Sanderson, J. (2018). Understanding the Artificial Intelligence Business Ecosystem. *IEEE Engineering Management Review*, *46*(4), 22–25. https://doi.org/10.1109/EMR.2018.2882430

Ross, C., & Swetlitz, I. (2018). *IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show.* https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf

Neff, G., & Nagy, P. (2016). Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication*.

Nissenbaum, H. (1994). Computing and accountability. *Communications of the ACM*, *37*(1), 72–80. https://doi.org/10.1145/175222.175228

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First paperback edition). Broadway Books.

Quan, X. I., & Sanderson, J. (2018). Understanding the Artificial Intelligence Business Ecosystem. *IEEE Engineering Management Review*, *46*(4), 22–25. https://doi.org/10.1109/EMR.2018.2882430

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach. Prentice Hall series in artificial intelligence*. Prentice Hall; London : Prentice-Hall International.

Simon, J., Wong, P.-H., & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1534

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701–717. https://doi.org/10.1006/ijhc.1999.0349

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. *SSRN*

Electronic Journal. Advance online publication. https://doi.org/10.2139/ssrn.3899991

Strahilevitz, L. J. (2008). Privacy versus antidiscrimination. *The University of Chicago Law Review*, *75*(1), 363–381.

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, *56*(4), 24–31. https://doi.org/10.1109/MSPEC.2019.8678513

Vallor, S., & Bekey, G. A. (2017). Artificial Intelligence and the Ethics of Self-learning Robots. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot ethics 2.0: New challenges in philosophy, law, and society* (pp. 338–353). Oxford University Press.

Veale, M., & Zuiderveen Borgesius, F. (2021). *Demystifying the Draft EU Artificial Intelligence Act*. Pre-print, July 2021. Version 1.1. https://doi.org/10.31235/osf.io/38p5f

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., & Petrov, S. (2020). *Measuring and Reducing Gendered Correlations in Pre-trained Models*. http://arxiv.org/pdf/2010.06032v2

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing Medical Imaging Data for Machine Learning. *Radiology*, *295*(1), 4–15. https://doi.org/10.1148/radiol.2020192224

Wolf, M. J., Miller, K. W [K. W.], & Grodzinsky, F. S. (2017). Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications. *The ORBIT Journal*, *1*(2), 1–12. https://doi.org/10.29297/orbit.v1i2.49

Yeung, K. (2019). Why Worry about Decision-Making by Machine? In K. Yeung & M. Lodge (Eds.), *Algorithmic regulation* (pp. 21–48).

Zweig, K. A., Wenzelburger, G., & Krafft, T. D. (2018). On Chances and Risks of Security Related Algorithmic Decision Making Systems. *European Journal for*

Security Research, *3*(2), 181–203. https://doi.org/10.1007/s41125-018-0031-2

10 Publication 4:

Reexamining computer ethics in light of AI systems and AI regulation

Authored by Mattis Jacobs and Judith Simon

Jacobs, M., & Simon, J. (2022). Reexamining computer ethics in light of AI systems and AI regulation. AI and Ethics. Advance online publication. https://doi.org/10.1007/s43681-022-00229-6

Citation style and bibliography harmonized

Reexamining computer ethics in light of AI systems and AI regulation

Mattis Jacobs, Judith Simon

Abstract

This article argues that the emergence of AI systems and AI regulation showcases developments that have significant implications for computer ethics and make it necessary to reexamine some key assumptions of the discipline. Focusing on design- and policy-oriented computer ethics, the article investigates new challenges and opportunities that occur in this context. The main challenges concern how an AI system's technical, social, political, and economic features can hinder a successful application of computer ethics. Yet, the article demonstrates that features of AI systems that potentially interfere with successfully applying some approaches to computer ethics are (often) only contingent, and that computer ethics can influence them. Furthermore, it shows how computer ethics can make use of how power manifests in an AI system's technical, social, political, and economic features to achieve its goals. Lastly, the article outlines new interdependencies between policy- and design-oriented computer ethics, manifesting as either conflicts or synergies.

1. Introduction

The emergence of Artificial Intelligence (AI) systems[1] leads to various new types of ethical issues. The relatively new discipline of AI ethics investigates these issues and develops approaches to address them. In this endeavor, AI ethics does not 'start from scratch' but builds on a rich and well-established body of literature on computer ethics. However, this article argues that computer ethics does not only provide concepts and methods for an ethical approach to designing and using AI systems. It makes the case that the emergence of AI systems and AI regulation also showcases developments that have significant implications for computer ethics and make it necessary to reexamine

[1] In this article, the term 'Artificial Intelligence' refers to Machine Learning-based Artificial Intelligence only. It does not refer to, for instance, knowledge-based "expert systems," in which problem-specific knowledge is formalized to enable rule-based decisions (cf. Russell & Norvig, 1995).

key assumptions of the discipline. Some challenges that the emergence of AI systems poses for individual approaches and methods of computer ethics have already been discussed in recent publications (Friedman et al., 2021; Umbrello, 2019; Umbrello & van de Poel, 2021). This article contributes to this discourse by discussing two further trends showcased by the emergence of AI systems and AI regulation. It raises the questions of which challenges they pose, which opportunities they provide, and how they affect the relationship among different approaches to computer ethics.

First, as James Moor stated in his seminal article "What Is Computer Ethics?" (Moor, 1985), policy-oriented computer ethics builds on the assumption that emerging computer technologies–such as AI systems–pose ethical issues primarily because they provide us with new capabilities. He argues that such new capabilities entail "new choices for actions" (Moor, 1985), which exist in a policy vacuum. That is, given these new choices for actions, often "no policies for [ethical] conduct" exist or "existing policies seem inadequate" (Moor, 1985), as the developments in computer technology outpace "ethical, [...] and legal developments" (Floridi & Sanders, 2002). The significance of Moor's concept of policy vacuum for computer ethics roots in that it serves as justification for the existence of computer ethics as a stand-alone discipline (Floridi & Sanders, 2001; Johnson, 1999; Maner, 1999; Moor, 2001) as well as that it establishes a core set of research questions for the field (Brey, 2000; Moor, 1985, 2001). However, unlike many other computer technologies, the emergence of AI systems led to calls for regulation that relatively quickly resulted in policy advancements. For instance, the European Commission recently proposed the Artificial Intelligence Act (AI Act), laying down comprehensive rules for AI systems (European Commission, 2021c). In light of the emergence of AI regulation, policy-oriented computer ethics, therefore, needs to address the question of which role it takes in highly regulated environments and how it takes into account and relates to existing policies.

Secondly, computer ethics and related disciplines discuss power primarily in terms of how technology affects "the way in which power is distributed and exercised in society" (Brey, 2008). Furthermore, to ensure that technology supports achieving an ethically sound distribution of power in society, various scholars call for stakeholder integration in design processes or a 'democratization of technology' (Friedman et al., 2021; Sclove, 1992; Slota, 2020; Zimmerman, 1995). However, as Friedman et al. (2021) note, computer ethics focusing on the design of computer systems often did not sufficiently consider power relationships among actors once they are involved in design processes. In the context of AI systems, such issues are especially prevalent. This is because AI systems are a prime example of computer systems consisting of various technical components which are usually developed and operated by relatively independent actors (Floridi & Sanders, 2004; Jacobs & Simon, 2022a). These include, among others, actors or groups of actors involved in data management and data preparation, model development, as well as deployment, use, and refinement of such systems (Jacobs & Simon, 2022a). Consequently, agency is highly distributed in the design of AI systems (Slota, 2020), and the question of how to account for power imbalances among actors involved in design processes deserves particular consideration (Jacobs et al., 2021).

Thus, the emergence of AI systems and AI regulation raises questions that computer ethics needs to address. This article reexamines computer ethics in light of the emergence of AI systems and AI regulation by investigating new challenges and opportunities. It does not aim at developing AI-specific solutions to discussed challenges but uses AI as an example to analyze how computer ethics needs to evolve in changing socio-technical environments. It focuses on policy- and design-oriented computer ethics, as these approaches to computer ethics are most clearly affected by the emergence of AI systems and AI regulation. Moreover, this article will demonstrate that novel interdependencies occur between the two approaches to computer ethics as a result.

The article proceeds as follows: Sect. 2 provides an overview of different approaches to computer ethics as well as the implications of the emergence of AI systems and AI regulation for these approaches. Furthermore, this section also outlines how computer ethics and AI ethics relate to each other. Section 3 discusses novel challenges arising in light of the emergence of AI systems and AI regulation, whereas Sect. 4 explores novel opportunities. Section 5 addresses new interdependencies between policy- and design-oriented computer ethics, manifesting as either conflicts or synergies. Lastly, Sect. 6 concludes by highlighting the key insights of this article and reflecting on the requirements for a productive integration of design- and policy-oriented computer ethics in light of these findings.

2. Computer ethics and the emergence of AI systems and AI regulation

According to van den Hoven (2008), “[c]omputer ethics is a form of applied or practical ethics [which] studies the moral questions that are associated with the development, application, and use of computers and computer science.” Computer ethics has developed over several decades, and perspectives of computer ethics have evolved significantly over time. While computer ethics can be traced back to Wiener’s cybernetics and information ethics (Wiener, 1961, 1989), the term itself was coined by Walter Maner and his computer ethics initiative in the mid-1970s (Bynum, 2008). Earlier publications focus primarily on practices relating to computer technology (especially its use) and, on a more abstract level, the challenges to existing ethical concepts (Bynum, 2008; Weizenbaum, 1976). Later, computer ethics began to also examine policies that guide actions enabled by computer technology (Moor, 1985), the professional conduct of computer specialists (ACM, 1992; Gotterbarn, 1991; Johnson & Miller, 2009), and the design of computer technology itself (Brey, 2000; Friedman et al., 2008; Nissenbaum, 2005). In line with the aim of this article, the remainder of this section focuses on policy- and design-oriented computer ethics in more detail. However, it first addresses the relationship between computer ethics and AI ethics to provide conceptual clarity.

2.1 Computer ethics and AI ethics

The rapid development and dissemination of AI systems in recent years has been “accompanied by constant calls for applied ethics” (Hagendorff, 2020). In response, AI ethics emerged and gained significant public and scholarly attention (Müller, 2020). While there is not necessarily a “categorical difference between computer ethics and the ethics of AI” (Stahl, 2022)–one can be understood as a subset of the other–the discourses in the two disciplines differ in some respects. Stahl (2022) identifies differences regarding, for instance, the scope, topics and issues, theoretical basis and referenced disciplines, solutions and mitigation, as well as importance and impact.

In its evolution, AI ethics did not customize the entirety of the methods and theories of computer ethics for the AI context. Rather, it focuses mainly on AI-specific issues. Yet, as AI systems are computer systems, the more general computer ethics remains highly relevant in the context of AI. It provides method and theory which can support understanding and addressing ethical issues of AI systems. However, as outlined in the introduction, the emergence of AI systems and AI regulation showcases developments that have significant implications for computer ethics which make it to reexamine key assumptions of the discipline.

The issues these developments pose for computer ethics are not necessarily unique to the AI context. For instance, just like AI systems, platform-ecosystems face increasing regulation (European Commission, 2020b, 2020c), and blockchain-based systems raise the question of who among the involved actors has the power to impose design decisions regarding the system's protocol (Jacobs et al., 2021; Walch, 2019a, 2019b). Thus, some of the challenges for computer ethics discussed in this article arise als o in other contexts. Yet, AI is an exceptional case to discuss and reflect on these developments, as many recent trends in the development of computer technology occur simultaneously in the context of AI and, therefore, can be examined in relation to each other.

Thus, the emergence of AI systems and AI regulation does not necessarily require developing a customized version of computer ethics for AI. Accordingly, this article attempts to reexamine (general) computer ethics in light of AI systems and AI regulation to identify challenges that these systems pose for selected approaches of the discipline.

2.2 Policy-oriented computer ethics

Moor (1985) holds the view that "computer ethics [emphasis in original] is the analysis of the nature and social impact of computer technology and the corresponding formulation and justification of policies for the ethical use of such technology." This reasoning is based on the observation that novel computer technologies "provide us with new capabilities [which] in turn give us new choices for actions" (Moor, 1985). Therefore, the emergence of new computer systems often results in situations "in which we do not have adequate policies in place to guide us" (Moor, 2005). Thus, according to Moor, computer ethics aims to develop coherent conceptual frameworks for understanding ethical problems involving computer technology and ultimately to replace such "policy vacuums with good policies supported by reasonable justifications" (Moor, 2001).

Addressing policy vacuums is especially important in computer ethics because the "logical malleability" of computer technology makes it a universal tool that enables human beings to do an "enormous number of new things" (Bynum, 2008). This vast field of application means that computer technology "can produce policy vacuums in larger quantities than other technologies" (Moor, 2001). This finding also applies to AI systems.

However, policy-makers increasingly push toward passing "legally binding regulations" addressing some of the ethical issues AI systems pose (Jobin et al., 2019). A prime example of this push is the EC's proposal for the AI Act. The AI Act is a policy proposal laying down harmonized rules on Artificial Intelligence in the European Union (European Commission, 2021c). The rules concern "the development, placement on the market and use of AI systems." Depending on the risk that a system poses, they include, for instance, "prohibitions and a conformity assessment system adapted from EU product safety law" (Veale & Zuiderveen Borgesius, 2021). Thus, once the AI Act goes into effect, AI systems in the EU are deployed in a highly regulated environment.

This does not mean that policy vacuums are no longer a concern of computer ethics. For instance, as Smuha et al. (2021) note, the AI Act's "list of prohibited practices seems heavily inspired by recent controversies." Therefore, future AI systems that enable not yet possible activities could exist in a policy vacuum again. Nevertheless, computer ethics will increasingly be applied in highly regulated contexts. Moving forward, computer ethics, therefore, needs to reflect on its role in contexts where there is no longer a policy vacuum.

2.3 Design-oriented computer ethics

Following the design turn in applied ethics, which had directed attention to the "design of institutions, infrastructure, and technology," also computer ethics began to address the design of computer technology itself (i.e., separate from the behavior of the developers and designers) (van den Hoven, 2008). Disclosive Computer Ethics (Brey, 2000, 2010; Introna, 2005) and Value Sensitive Design (Friedman et al., 2008; Friedman & Hendry, 2019) are indicative approaches for the design turn in computer ethics. Both approaches argue that problems of computer ethics can be solved not only by developing policy regulating practices relating to computer systems (e.g., their use) but also by accounting for ethical values and principles in the design of computer technology. To achieve an alignment of technology design and ethical values, computer ethics accounts for values as well as “norms, practices, and incentives, perhaps originating from different stakeholders” (Friedman & Hendry, 2019).

The more analytic Disclosive Computer Ethics focuses “on morally opaque practices” (Brey, 2010) and the “moral deciphering of computer technology” or, more specifically, its “design features.” It concerns the exposure of opaque moral features (or “embedded normativity”) of computer technology (Brey, 2000). The more constructive Value Sensitive Design argues that moral features in the design of (computer) technology can not only be analyzed ex-post but can already be accounted for in design processes. It “provides theory, method, and practice to account for human values in a principled and systematic manner throughout the technical design process” (Friedman & Hendry, 2019).

However, AI systems consist of several components, such as training data, an algorithm that infers decision rules from data based on a learning method, an algorithm that classifies cases based on the learned decision rules, and some form of end-user application that uses these classifications and translates them into decisions (Keller et al., 2018; Krafft et al., 2020; Yeung, 2019). In many instances, the required system components are developed and/or controlled not by one but by several actors who specialize in one or a few tasks in the development of AI systems (Dhinakaran, 2020). Furthermore, not only the engineers and data scientists directly involved in the development of Machine Learning capabilities can influence design decisions, but also “their [respective] managers, product designers, clients, executives, and others” (Boyd, 2022).

Hence, AI systems need to be understood as “a complex network” of technical and non-technical components in which individual designers often lack the capacity to steer or control the design of the system at large (Slota, 2020). As Barocas & Selbst (2016), Danks & London (2017), and others show, ethically problematic features of AI systems can have their roots in tasks performed by many of the involved actors, such as data management and data preparation, model development, as well as deployment, use, and refinement of AI systems. Consequently, the distribution of agency regarding design decisions among the involved actors poses challenges for addressing threats to the realization of ethical values, the consideration of ethical principles, and fundamental rights such systems can pose (Jacobs & Simon, 2022a). Prominent advocates of design-oriented computer ethics, such as Friedman et al. (Friedman et al., 2021), acknowledge this as one of the 'grand challenges' that the discipline is facing today. They note that many Value Sensitive Design projects have assumed that the "practices, organizational policies, or legal frameworks in place will support 'doing the right thing,' without needing to be explicit about the role and importance of power relationships" among the involved actors.

Considering the emergence of more vast and complex socio-technical systems such as AI systems, these power relationships gain importance, as they affect how computer ethics can engage in designing such systems. In the case of such systems, computer ethics needs to account for how agency is distributed among the actors involved in AI systems and to what extent involved actors have the power to realize design decisions. This article adopts a definition of power focusing on outcomes,[2] according to which power is the "ability of agents" to "realize a certain outcome" or "bring about certain [...] state of affairs" (Brey, 2008; see also Dowding, 1996). The emergence of more vast and complex socio-technical systems such as AI systems raises questions for computer ethics that go beyond how technology affects how power is distributed in society and which societal actors should take part in designing technical artifacts. It also raises the question of how power manifests in the broader social, economic, and political features of such systems (cf. Sattarov, 2019), as these features co-determine the ability of actors involved in the design process to ultimately impact design decisions. Computer ethics needs to address the question of which of the actors involved in the development and operation of AI systems have (and should have) the ability to realize specific design decisions.[3]

3. Novel challenges for computer ethics

Based on the explanatory notes in Sect. 2, the following paragraphs exemplify how the emergence of AI systems and AI regulation challenge computer ethics in practice.

[2] This conception of power to realize outcomes is contrasted with a conception of power in terms of power over other actors (Brey, 2008).

[3] As noted in the introduction, this article does not suggest that power related issues have not been discussed in computer ethics generally. The reasoning of this article focuses on power-related issues among actors who are involved in development processes in different ways specifically.

Concerning AI systems, biased decision models that unfairly discriminate against groups or individuals are a widely discussed ethical issue. Such bias can be caused by various factors, such as biased training data or algorithms (Barocas & Selbst, 2016; Danks & London, 2017). However, in many cases, some actors involved in developing AI systems can account for and mitigate such biases to prevent that biased training data or a biased algorithm lead to problematic outcomes. For instance, as Danks & London (2017) note, algorithmic processing can be used to "offset or correct for" biased training data, or the end-user application in which an AI system is embedded can be set up in a way that it does "not take action solely on the basis of the algorithm output" in cases where a biased output is to be expected. This way, developers can attempt to "develop a system that is overall unbiased, even though different components each exhibit [...] bias."[4] To account for and react to biases, developers of decision models or end-user applications need information on properties of the training data or the decision model, respectively (Jacobs & Simon, 2022a). However, due to, for instance, business interests, actors involved in managing and preparing training data or developing decision models can decide against providing access to this information (Keller et al., 2018), even at the cost of negatively affecting the ability of developers and users to account for and react to this ethical issue.

As design specifications in both novel regulation and regulatory proposals demonstrate, policy-makers are further actors that can impact the ability of actors involved in socio-technical systems to account for specific ethical values and principles in design. For instance, the proposal for the AI Act prohibits the deployment of certain types of AI systems. It prescribes technical and non-technical requirements for the (legal) use of AI systems by the threat of penalties. Thereby, it incentivizes certain design decisions while disincentivizing others. In the AI Act, obligations concern, for instance, the establishment of quality management systems, the provision of technical documentation, or ensuring data governance in accordance with specified standards (European Commission, 2021c). Often, such obligations reflect specific ethical principles or values, such as privacy, fairness, or transparency. Moreover, as promoting one value or principle often comes at the expense of another, they also reflect value tradeoffs. For instance, as Sect. 5 discusses in more detail, privacy regulations can hamper bias mitigation strategies that require integrating more data

[4] For an overview on bias detection and mitigation strategies see also Fu et al. (2020).

(Jobin et al., 2019). Furthermore, values like fairness can be defined in various conflicting ways. Thus, requiring an AI system to make fair decisions according to one definition of fairness makes it impossible to achieve fair decisions according to a conflicting definition of fairness (Binns, 2018). Therefore, such regulatory interventions can obstruct or compel design decisions that promote or demote the realization of specific values (Jacobs et al., 2021). Consequently, they can reduce the developers' scope for design and hamper their ability to negotiate and account for values themselves or in accordance with further stakeholders.

Such limited agency of developers regarding design decisions poses new conceptual and practical challenges for design- and policy-oriented computer ethics. As technical, social, economic, and political features of a socio-technical system like an AI system can cut back on the involved actors' ability to design technical components in accordance with ethical values and principles, design-oriented computer ethics needs to consider not only what designers ought to do and how technology should be designed. It also needs to address the questions of what individual developers have the ability to do, what constraints there are for design decisions, and which actors set these constraints. For more analytical approaches to design-oriented computer ethics, such as Disclosive Computer Ethics, the question arises which of the involved actors has the ability to address problematic ethical features of computer systems that are integrated into larger socio-technical systems once these features have been disclosed. For more constructive approaches to design-oriented computer ethics, such as Value Sensitive Design, the question arises which actors involved in a socio-technical system can assert design decisions that align with specific ethical principles or values and can, therefore, successfully apply these approaches. Conversely, they also have to engage with the question of which actors lack the ability to apply them successfully and how they can change this circumstance (Jacobs & Simon, 2022a).

Policy-oriented computer ethics also faces challenges in light of the complex actor constellations in AI systems. In the process of making policy for the ethical use of computer technology, it needs to take the ability of actors to achieve certain outcomes into account. This is because if policy-makers do not outright ban specific applications but assign obligations to their development, deployment, or use, these obligations need to be assigned to some role or actor. Yet, if policy-makers assign obligations to actors incapable of fulfilling them, these obligations will not achieve the intended results. While this may seem trivial in theory, it leads to major challenges in practice. For instance, if policy-oriented computer ethics seeks to ensure that actors involved in AI systems warrant that potential bias in training data does not lead to biased decisions that harm individuals, it is challenging to determine which involved actors can or should be addressed: actors in charge of data collection and management (to ensure that there is no bias in the training data), actors in model development (to ensure that compensatory bias is applied so that decisions are unbiased), operators (to question decisions and not rely on them in cases that decisions might be biased), or providers (Danks & London, 2017; Jacobs & Simon, 2022a).[5]

Thus, the rise of more vast and complex socio-technical systems such as AI systems forces policy-oriented computer ethics to determine not only what ethical practices relating to computer technology are (Moor, 1985) but also which actors have the ability to engage in these practices and which actors the respective obligations should be assigned to. To ensure the intended effects of policy measures, it is crucial to account for the involved actors' power to achieve specific outcomes.

4. Novel opportunities for computer ethics

As the discussion of challenges in Sect. 3 shows, computer ethics needs to account for the complex actor constellations in socio-technical systems such as AI systems and consider how power manifests in their broader technical, social, economic, and political features. However, these features should not be perceived as unchangeable or as (only) a hindrance to computer ethics. The way that the technical, social, political, and economic features of socio-technical systems determine the power of involved actors to shape the design of computer technology is contingent. It can be influenced in a variety of ways (Jacobs et al., 2021). Enabling and stimulating ethical reflection and conduct by impacting these features of socio-technical systems should thus be seen as a field of activity for computer ethics. Furthermore, computer ethics can make use of how power manifests in socio-technical systems to achieve its goals.

The new opportunities for design-oriented computer ethics are twofold. First, it can propose design features for technical components that co-determine the ability of

[5] Section 4 outlines how this challenge manifested in the development of the AI Act and how the current proposal for the AI Act addresses it.

actors involved in the socio-technical system to achieve specific outcomes. While this approach does not directly facilitate accounting for a specific ethical value or principle within the given system, it enables actors to apply methods of computer ethics. Second, acknowledging differences in the ability to assert design decisions among actors involved in a socio-technical system can help to identify powerful actors. These actors can then be encouraged to enforce compliance with specific ethical values or principles in the socio-technical system at large.

The first approach, that is, engaging with the technical, social, political, and economic features of a socio-technical system to determine the ability of actors involved in the system to achieve specific goals of computer ethics, can be achieved, for instance, by aiming for transparency in the system's design. Section 3 described how a lack of information on, for instance, training data or properties of a decision model can hamper efforts to mitigate bias in AI systems. Conversely, a higher degree of transparency on properties of training data and the decision model can support actors in accounting for these properties and compensate for bias. A more transparent design allows for "a broader conversation about the values, operation, and limitations" of an AI system and can thereby foster the ability of involved actors to account for ethical values and principles in the system's design (Slota, 2020). Yet, if achieving greater transparency conflicts with other ethical values (or business interests), it is necessary to weigh these values (or interests) against each other (Granka, 2010; Mittelstadt et al., 2016).

The second approach, that is, focusing on powerful actors to ensure that specific ethical values or principles are accounted for, can be demonstrated by the use of data access control for protecting sensitive data. Actors developing machine-learning-based AI systems often strive for ever more data to enhance the respective system's quality and accuracy (Keller et al., 2018). However, as Yeung (2019) notes, this striving for ever more data can be ethically problematic. Individuals can have "a legitimate interest in not being evaluated and assessed" based on information that is "morally and/or causally irrelevant" to the decision," even if this information "may have a very high degree of predictive value (i.e., statistical relevance)." In contexts where data are not widely available and individual actors control specific information exclusively, these actors can use their position by not granting access to specific types of sensitive information. Thereby, these actors can prevent this information from being used to train a decision model or make decisions, even if they are not directly involved in either of these activities. This is possible, for instance, where personal tracking devices generate otherwise unavailable data.[6]

Thus, if design-oriented computer ethics is applied by actors in a dominant position in a socio-technical system, these actors can not only affect the design of technical components or applications which they are designing. To a varying degree, they can also shape the broader socio-technical system by co-determining if (and if so, how) values are accounted for in the system at large (Jacobs et al., 2021).

However, the two approaches can be in conflict with one another. This is because attempts by an actor to shape a socio-technical system as a whole (including technical components and applications that are developed and operated by other actors) require that this actor has a certain assertiveness regarding design decisions and other actors do not. For instance, because sensitive information can be used to identify or mitigate bias (cf. Fu et al., 2020), not granting access to sensitive information can get in the way of efforts to identify or mitigate bias in AI systems. Thus, protecting privacy can conflict with achieving unbiased decisions (Jobin et al., 2019). Moreover, if design-oriented computer ethics uses an actor's dominant position within a socio-technical system to promote a specific ethical value, this can hinder efforts of other actors to negotiate and account for different values. Consequently, conflicts can arise between applying design-oriented computer ethics to determine design decisions in accordance with ethical values or applying it to enable further actors to engage in ethical considerations in design processes.

Moreover, new opportunities arise not only for design-oriented computer ethics but also for policy-oriented computer ethics. As design-oriented computer ethics, policy-oriented computer ethics can acknowledge and make use of existing features of a socio-technical system to achieve its goals or attempt to influence them. It can take advantage of how power manifests in a socio-technical system's technical, social, political, and economic features by targeting and assigning obligations specifically to actors who have a powerful position because of these features. The proposal for the AI Act provides a prime example of this approach. Earlier whitepapers on AI assign

[6] Please note that in many other cases, data is more widely available. Often, the same information can be provided by different data controllers (Christl, 2017) or inferred by using more publicly available proxy data Danks and London (2017). Consequently, other power relations exist in these cases.

obligations to ensure specific properties of socio-technical systems are met to actors who are “best placed” to address them. To illustrate this approach, the whitepaper clarifies that, for example, “while the developers of AI may be best placed to address risks arising from the development phase, their ability to control risks during the use phase may be more limited. In that case, the deployer should be subject to the relevant obligation” (European Commission, 2020a). The AI Act, instead, assigns most obligations to providers[7] of AI systems. It uses the providers’ position—characterized by providing market access—to ensure that other actors involved in the respective AI system ensure adequate data governance, provide technical documentation, or establish a quality management system (European Commission, 2021c). In doing so, the AI Act avoids the difficulties it would have faced if it did not delegate these tasks to providers, such as the need to engage in the micromanagement of assigning obligations according to the capabilities of individual actors involved in an AI system (Jacobs & Simon, 2022a).

Second, policy-oriented computer ethics can advocate policies that change how power manifests in a socio-technical system’s technical, social, political, and economic features. For instance, the “AI Act proposes a new, central database, managed by the Commission, for the registration of ‘stand-alone’ high-risk AI systems” to help actors such as the regulatory authorities, civil society, or journalists to “uncover illicit AI” (Veale & Zuiderveen Borgesius, 2021). The proposed database aims at “enhanced oversight by the public authorities and the society” of high-risk AI systems (European Commission, 2021b). Thus, this proposal fosters an infrastructure enabling various groups of actors to engage in discourses on the risks that AI systems pose for the realization of ethical values, the consideration of ethical principles, and fundamental rights. Furthermore, it challenges a status quo in which some actors groups are often excluded from these discourses.

[7] The AI Act defines providers as “a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed with a view to placing it on the market or putting it into service under its own name or trademark, whether for payment or free of charge” (European Commission, 2021c).

5. Conflicts and synergies

Based on the findings in the previous sections, it is evident that policy- and design-oriented computer ethics can engage with the same actors, computer systems, value conflicts, or–more generally–state of affairs. Both approaches can be applied to target specific actors involved in a socio-technical system to encourage or coerce them to ensure that one particular ethical value or principle is accounted for in the system at large. Moreover, both approaches can be applied to influence how the technical, social, political, and economic features of a socio-technical system co-determine the distribution of power among the actors involved in the system.

This raises the question of how the two approach

hes to computer ethics relate to one another. Brey (2000) argues that design-oriented computer ethics is complementary to “mainstream” or policy-oriented computer ethics. Yet, this assessment needs to be re-evaluated in light of the findings of the previous sections. This section makes the case that while these two approaches can be complementary (i.e., they can create synergies), they can also be in conflict with one another.

If design-oriented computer ethics is applied in contexts where policy constrains design decisions, developers and design-oriented computer ethicists need to take this circumstance into account. As outlined above, policies can affect the application of design-oriented computer ethics in two ways. On the one hand, they can affect the consideration of specific values in design decisions. On the other hand, they can affect the ability of actors involved in a socio-technical system to influence design decisions and thus shape technology in line with ethical values and principles. This can lead to conflicts if either the respective approaches to computer ethics promote conflicting values (or operationalizations of values) or if one approach aims at enhancing the ability of specific actors to achieve their goals in a way that counteracts the other approach.

For instance, as Jobin et al. (2019) note, “the need for ever-larger, more diverse datasets to ‘unbias’ AI might conflict with the requirement to give individuals increased control over their data and its use to respect their privacy and autonomy.” This can result in conflicts between design- and policy-oriented computer ethics. The European General Data Protection Regulation (GDPR), a regulation that primarily aims at enhancing data protection rights of individuals and thereby strengthening their fundamental rights in the digital age (European Commission, 2016), can conflict with approaches of design-oriented computer ethics aiming at mitigation bias in AI systems. Article 10(5) of the AI Act specifically addresses this issue and defines an exemption to the GDPR, which allows providers to "process special categories of data" according to Article 9(1) of the GDPR "to the extent that it is strictly necessary for the purposes of ensuring bias monitoring, detection and correction in relation to the high-risk AI systems" (European Commission, 2021c; see also Smuha et al., 2021). This exception in the data protection guidelines is needed to prevent the GDPR from obstructing bias mitigation strategies. Thus, depending on whether policy-oriented computer ethics promotes data protection regulation as initially defined in the GDPR or exceptions to data protection regulations as proposed by the AI Act, it conflicts with or complements such approaches to bias mitigation.

Moreover, design-oriented computer ethics can also counteract policy-oriented computer ethics. For instance, there is an ongoing debate if using AI to manipulate (or 'nudge') individuals into making choices for benign purposes, such as acting environmentally friendly, is ethically acceptable or not (Thaler & Sunstein, 2008; Yeung, 2019). Policy-oriented computer ethicists might conclude that such manipulation is ethically not justifiable and propose policies that prohibit it. However, defining practices that constitute ethically not acceptable manipulation is challenging. For example, the AI Act addresses this issue by prohibiting AI systems that deploy "subliminal techniques" (European Commission, 2021c) or exploit specific types of vulnerabilities linked to, for instance, "age, physical or mental disability" (European Commission, 2021c). However, as Smuha et al. (2021) note, this approach is "under-protective, as it only applies to the exploitation of a limited set of vulnerabilities" and leaves "the door open to many non-subliminal manipulative AI practices." Therefore, if design-oriented computer ethicists assume that manipulation for specific benign purposes is legitimate, they could exploit such legal loopholes by customizing the design of manipulative AI systems to evade regulation. In such a case, an AI system is not situated in a policy vacuum because it enabled novel actions that policy-makers did not yet consider. Instead, placing the system outside the scope of existing policy is the intention behind the respective design decisions.

Yet, as stated above, policy- and design-oriented computer ethics can also complement each other. As policies co-determine how a socio-technical system's technical, social, political, and economic features influence the ability of involved actors to assert design decisions, policy-oriented computer ethics can enable developers to apply approaches to account for ethical values and principles in the design process. For instance, individual powerful actors in socio-technical systems often can prevent developers from accounting for ethical values and principles in technical design if this conflicts with their commercial interests. Here, policy-oriented computer ethics can promote regulation that establishes a threat of fines for not ensureing that technology design accounts for specific (operationalizations of) values. In doing so, it can change the cost-/benefit-analysis of these actors and soften or end the resistance to design decisions in accordance with specific ethical values.

Conversely, the design of technical components of socio-technical systems also co-determines how well the respective socio-technical system can be regulated. For instance, in the case of AI systems, designing systems more transparently and providing explanations for how output is generated allows identifying problematic uses. In turn, this enables the "formulation and justification of policies for the ethical use of such technology" (Moor, 1985).

6. Conclusion

This article reexamines foundational assumptions of computer ethics in light of the emergence of AI systems and AI regulation. It outlines both challenges and highlights opportunities arising in this context. The main challenges concern how a socio-technical system's technical, social, political, and economic features can hinder a successful application of policy- and design-oriented computer ethics. Furthermore, the article underlines that powerful actors in socio-technical systems can intentionally influence these features to co-determine the ability of other actors involved in the socio-technical system to achieve specific outcomes. With advancing regulation, AI systems are often no longer deployed in policy vacuums, suggesting that policy-makers become such powerful actors. Thus, computer ethics will increasingly need to account for them as such in the future.

However, as mentioned before, this article argues that the emergence of AI systems and AI regulation does not only exemplify new challenges or hindrances for computer ethics. They also present new opportunities. Indeed, features of AI systems that potentially hinder a successful application of approaches to computer ethics are (often) only contingent, and computer ethics can influence them. Doing so can enable actors involved in designing and operating AI systems to account for ethical values and principles in the system's design and use. Furthermore, computer ethics can acknowledge and make use of how power manifests in the technical, social, political, and economic features of AI systems. It can use the powerful position of specific actors in AI systems to assert how ethical values and principles are being accounted for in design decisions or other practices relating to the respective AI system.

Furthermore, the emergence of AI systems and AI regulation showcases novel interdependencies between policy- and design-oriented computer ethics. These interdependencies manifest as either conflicts or synergies. Policy- and design-oriented computer ethics have been mainly discussed as being complementary in pertinent literature (Brey, 2000; Friedman, Smith, et al., 2006; Miller et al., 2007). However, this article shows that the two approaches can also be at odds with one another. Therefore, computer ethicists should engage with the question of if pursuing certain goals potentially has unintended effects on applying design- or policy-oriented computer ethics elsewhere in a socio-technical system that can lead to such conflicts. Further research should investigate ways to systematically avoid or resolve such conflicts (where they were not consciously caused) and establish complementarity.

If computer ethics takes the developments showcased by the emergence of AI systems and AI regulation into account and adapts accordingly, its methods and concepts become more applicable in the context of AI. This does not only provide new possibilities to computer ethics but also to AI ethics. Computer ethics profits by that it can more easily and effectively apply its methods and concepts in discourses related to the ethical issues of AI. This improved applicability might also exist in regard to other systems which share features such as complex constellations of involved actors, severe power imbalances, or a high degree of regulation with AI systems. AI ethics, on the other hand, profits by that it can incorporate methods of computer ethics more easily and, thereby, augment the methodological and conceptual toolkit available to it.

Lastly, there are two crucial limitations to this article. First, this article focused on design- and policy-oriented computer ethics specifically. However, as noted in Sect. 2, there are further approaches to computer ethics. Presumably, some of these approaches, such as professional ethics, are also affected by the developments discussed in this article. Further research should, therefore, examine the challenges and opportunities that arise due to the emergence of AI systems and AI regulation for these other approaches to computer ethics. Second, while the emergence of AI systems and AI regulation is a prime example to showcase the developments discussed in this article, they are not unique to AI. Policy advancements such as the European Digital Services Act (European Commission, 2020b) and the European Digital Markets Act (European Commission, 2020c) call into question the existence of a policy vacuum in relation to other computer technologies. Furthermore, other emerging technologies, such as blockchain technology, are also exhibiting power struggles among actors involved in design processes concerning the advancement of the respective system (Jacobs et al., 2021; Walch, 2019a, 2019b; Werbach, 2018). Thus, while the challenges and opportunities discussed in this article are well illustrated by the emergence of AI systems and AI regulation, they are similarly prevalent in other contexts—and offer a rich field of study for future research.

Acknowledgments

Mattis Jacobs received funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2002/1 "Science of Intelligence"—project number 390523135.

Open Access funding enabled and organized by Projekt DEAL. The authors have no relevant financial or non-financial interests to disclose.

Publication's References

ACM (1992). ACM code of ethics and professional conduct. *Communications of the ACM*, *35*(5), 94–99.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, *81*, 1–11.

Brey, P. (2008). The Technological Construction of Social Power. *Social Epistemology*, *22*(1), 71–95. https://doi.org/10.1080/02691720701773551

Brey, P. (2010). Values in Technology and Disclosive Computer Ethics. In L. Floridi (Ed.), *The Cambridge Handbook of Information and Computer Ethics*.

Boyd, K. (2022). Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development. In *ICPS, FAccT 2022: 2022 5th ACM Conference on Fairness, Accountability, and Transparency : June 21-24, 2022, Seoul, South Korea* (pp. 2069–2082). The Association for Computing Machinery. https://doi.org/10.1145/3531146.3534626

Bynum, T. W. (2008). Milestones in the History of Information and Computer Ethics. In K. E. Himma & H. T. Tavani (Eds.), *The Handbook of Information and Computer Ethics* (pp. 25–48). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470281819.ch2

Christl, W. (2017). Corporate surveillance in everyday life: How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions. Cracked Labs.

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. In F. Bacchus & C. Sierra (Eds.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4691–4697). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2017/654

Dhinakaran, A. (2020). *The AI Ecosystem is a MESS: Why is it impossible to understand what AI companies really do?* Towards Data Science. https://towardsdatascience.com/the-ai-ecosystem-is-a-mess-c46bdfbf43e4

Dowding, K. M. (1996). *Power. Concepts in the social sciences series*. Open University Press.

European Commission. (2016). *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

European Commission. (2020a). *On Artificial Intelligence - A European approach to excellence and trust: Whitepaper*.

https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission. (2020b). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en

European Commission. (2020c). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on contestable and fair markets in the digital sector (Digital Markets Act)*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0842&from=en

European Commission. (2021b). *COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT Accompanying the Proposal for a Regulation of the European Parliament and of the Council LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. https://eur-lex.europa.eu/resource.html?uri=cellar:0694be88-a373-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

Floridi, L., & Sanders, J. W. (2001). Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*, *3*(1), 55–66. https://doi.org/10.1023/A:1011440125207

Floridi, L., & Sanders, J. W. (2002). Mapping the foundationalist debate in computer ethics. *Ethics and Information Technology*, *4*(1), 1–9. https://doi.org/10.1023/A:1015209807065

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Friedman, B., Harbers, M., Hendry, D. G., van den Hoven, J., Jonker, C., & Logler, N. (2021). Eight grand challenges for value sensitive design from the 2016 Lorentz workshop. *Ethics and Information Technology*, *23*(1), 5–16. https://doi.org/10.1007/s10676-021-09586-y

Friedman, B., & Hendry, D. (2019). *Value sensitive design: Shaping technology with moral imagination*. The MIT Press.

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value Sensitive Design and Information Systems. In K. E. Himma & H. T. Tavani (Eds.), *The Handbook of Information and Computer Ethics* (pp. 69–101). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470281819.ch4

Friedman, B., Smith, I., H. Kahn, P., Consolvo, S., & Selawski, J. (2006). Development of a Privacy Addendum for Open Source Licenses: Value Sensitive Design in Industry. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, P. Dourish, & A. Friday (Eds.), *Lecture Notes in Computer Science. UbiComp 2006: Ubiquitous Computing* (Vol. 4206, pp. 194–211). Springer Berlin Heidelberg. https://doi.org/10.1007/11853565_12

Fu, R., Huang, Y., & Singh, P. V. (2020). Artificial Intelligence and Algorithmic Bias: Source, Detection, Mitigation, and Implications. In C. Druehl, W. Elmaghraby, D. Shier, & H. J. Greenberg (Eds.), *Pushing the Boundaries: Frontiers in Impactful OR/OM Research* (Vol. 65, pp. 39–63). INFORMS. https://doi.org/10.1287/educ.2020.0215

Gotterbarn, D. (1991). Computer ethics: responsibility regained. *National Forum: The Phi Beta Kappa Journal*, 26–31.

Granka, L. A. (2010). The Politics of Search: A Decade Retrospective. *The Information Society*, *26*(5), 364–374. https://doi.org/10.1080/01972243.2010.511560

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

Introna, L. D. (2005). Disclosive Ethics and Information Technology: Disclosing Facial Recognition Systems. *Ethics and Information Technology*, *7*(2), 75–86. https://doi.org/10.1007/s10676-005-4583-2

Jacobs, M., Kurtz, C., Simon, J., & Böhmann, T. (2021). Value Sensitive Design and power in socio-technical ecosystems. *Internet Policy Review*, *10*(3). https://doi.org/10.14763/2021.3.1580

Jacobs, M., & Simon, J. (2022). Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed by the European Commission. *Digital Society*, *1*(1), Article 6. https://doi.org/10.1007/s44206-022-00009-z

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

Johnson, D. G. (1999). Sorting out the uniqueness of computer-ethical issues.

Johnson, D. G., & Miller, K. W. (2009). *Computer ethics: Analyzing information technology* (4. ed.). Pearson Education Intern.

Krafft, T. D., Zweig, K. A., & König, P. D. (2020). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance*, *104*, 119–136. https://doi.org/10.1111/rego.12369

Keller, J. R., Chauvet, L., Fawcett, J., & Thereaux, O. (2018). *The role of data in AI business models*. Open Data Institute. https://theodi.org/wp-content/uploads/2018/04/376886336-The-role-of-data-in-AI-business-models.pdf

Maner, W. (1999). Is computer ethics unique?

Miller, J. K., Friedman, B., & Jancke, G. (2007). Value tensions in design. In T. Gross & K. Inkpen (Eds.), *Proceedings of the 2007 international ACM conference on Conference on supporting group work - GROUP '07* (p. 281). ACM Press. https://doi.org/10.1145/1316624.1316668

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

Moor, J. H. (1985). What Is Computer Ethics? *Metaphilosophy*, *16*(4), 266–275. https://doi.org/10.1111/j.1467-9973.1985.tb00173.x

Moor, J. H. (2001). The future of computer ethics: You ain't seen nothin' yet! *Ethics and Information Technology*, *3*(2), 89–91. https://doi.org/10.1023/A:1011881522593

Moor, J. H. (2005). Why We Need Better Ethics for Emerging Technologies. *Ethics and Information Technology*, *7*(3), 111–119. https://doi.org/10.1007/s10676-006-0008-0

Müller, V. C. (2020). *Ethics of Artificial Intelligence and Robotics*. https://plato.stanford.edu/entries/ethics-ai/

Nissenbaum, H. (2005). Values in Technical Design. In C. Mitcham (Ed.), *Encyclopedia of science, technology, and ethics* (pp. lxvi–lxx). Macmillan Reference USA.

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. *Prentice Hall series in artificial intelligence*. Prentice Hall; London : Prentice-Hall International.

Sattarov, F. (2019). *Power and technology: A philosophical and ethical analysis*. Rowman et Littlefield.

Sclove, R. E. (1992). The Nuts and Bolts of Democracy: Democratic Theory and Technological Design. In L. Winner (Ed.), *Democracy in a Technological Society* (pp. 139–157). Springer Netherlands. https://doi.org/10.1007/978-94-017-1219-4_9

Slota, S. C. (2020). Designing Across Distributed Agency: Values, participatory design and building socially responsible AI. *Good Systems-Published Research*.

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.3899991

Stahl, B. C. (2022). From computer ethics and the ethics of AI towards an ethics of digital ecosystems. *AI and Ethics*, *2*(1), 65–77. https://doi.org/10.1007/s43681-021-00080-1

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Umbrello, S. (2019). Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach. *Big Data and Cognitive Computing*, *3*(1), 5. https://doi.org/10.3390/bdcc3010005

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, *1*(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3

van den Hoven, J. (2008). Moral Methodology and Information Technology. In K. E. Himma & H. T. Tavani (Eds.), *The Handbook of Information and Computer Ethics* (pp. 49–67). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470281819.ch3

Veale, M., & Zuiderveen Borgesius, F. (2021). *Demystifying the Draft EU Artificial Intelligence Act*. Pre-print, July 2021. Version 1.1. https://doi.org/10.31235/osf.io/38p5f

Walch, A. (2019a). Deconstructing 'Decentralization': Exploring the Core Claim of Crypto Systems. *SSRN Electronic Journal*.

Walch, A. (2019b). In Code(rs) We Trust: Software Developers as Fiduciaries in Public Blockchains. In I. Lianos, P. Hacker, S. Eich, & G. Dimitropoulos (Eds.), *Regulating Blockchain: Techno-Social and Legal Challanges* (pp. 58–81). Oxford University Press.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. Freeman.

Werbach, K. (2018). *The blockchain and the new architecture of trust*. *Information policy series*. The MIT Press.

Wiener, N. (1961). *Cybernetics or control and communication in the animal and the machine* (2. ed., 14. print). MIT Press.

Wiener, N. (1989). *The human use of human beings: Cybernetics and society*. Free Association.

Yeung, K. (2019). Why Worry about Decision-Making by Machine? In K. Yeung & M. Lodge (Eds.), *Algorithmic regulation* (pp. 21–48).

Zimmerman, A. D. (1995). Toward a more democratic ethic of technological governance. *Science, Technology, & Human Values*, *20*(1), 86–107.

References

Abdul-Rahman, A., & Hailes, S. (1998). A distributed trust model. In *Proceedings of the 1997 workshop on New security paradigms*. Symposium conducted at the meeting of ACM.

ACM (1992). ACM code of ethics and professional conduct. *Communications of the ACM*, *35*(5), 94–99.

Alsindi, W. Z., & Lotti, L. (2021). Mining. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1551

Anderson, J., & Mariniello, M. (2021). *Regulating big tech: the Digital Markets Act*. bruegel.org. https://www.bruegel.org/2021/02/regulating-big-tech-the-digital-markets-act/

Antonopoulos, A. (2014). *Bitcoin security model: Trust by computation*. O'Reilly. Radar. http://radar.oreilly.com/2014/02/bitcoin-security-model-trust-by-computation.html

Antonopoulos, A. (2017). *Mastering Bitcoin: Programming the open Blockchain* (Second edition). O'Reilly Media.

Apple. (2019). *Turn Location Services and GPS on or off on your iPhone, iPad, or iPod touch*. https://support.apple.com/en-us/HT207092

Apple. (2020). *App Tracking Transparency: Request user authorization to access app-related data for tracking the user or the device*. https://developer.apple.com/documentation/apptrackingtransparency

Arbel, T. (2019). *Researchers say Amazon face-detection technology shows bias*. https://abcnews.go.com/Technology/wireStory/researchers-amazon-face-detection-technology-shows-bias-60630589?cid=social_twitter_abcn

Atzori, M. (2015). Blockchain Technology and Decentralized Governance: Is the State Still Necessary? *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.2709713

Atzori, M., & Ulieru, M. (2017). Architecting the eSociety on Blockchain: A Provocation to Human Nature. *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.2999715

Ausloos, J., & Veale, M. (2020). *Researching with Data Rights*. https://doi.org/10.26116/techreg.2020.010

Baier, A. (1986). Trust and Antitrust. *Ethics*, *96*(2), 231–260. https://doi.org/10.1086/292745

Baldwin, C. Y., & Woodard, C. J. (2009). The architecture of platforms: A unified view. In A. Gawer (Ed.), *Platforms, Markets and Innovation* (pp. 19–44). Edward Elgar Publishing.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*, 671.

Bauer, J. M., & Herder, P. M. (2008). Designing Socio-Technical Systems. In *Handbook of the philosophy of science: [9]. Handbook of philosophy of technological sciences* (pp. 601–630). Elsevier. https://doi.org/10.1016/B978-0-444-51667-1.50026-4

Beck, R., Czepluch, J. S., Lollike, N., & Malone, S. (2016). Blockchain-the Gateway to Trust-Free Cryptographic Transactions. In *ECIS*.

Becker, M., & Bodó, B. (2021). Trust in blockchain-based systems. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1555

Bietti, E. (2021). From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics. *Journal of Social Computing*, *2*(3), 266–283. https://doi.org/10.23919/JSC.2021.0031

Binance. (2019). *Binance Will Delist BCHSV*. https://binance.zendesk.com/hc/en-us/articles/360026666152

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research*, *81*, 1–11.

Bishop, C. M. (2006). *Pattern recognition and machine learning. Information Science and Statistics*. Springer Science+Business Media LLC. https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/

Bodó, B., Brekke, J. K., & Hoepman, J.-H. (2021). Decentralisation: a multidisciplinary perspective. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1563

Bolotaeva, O. S., Stepanova, A. A., & Alekseeva, S. S. (2019). The Legal Nature of Cryptocurrency. *IOP Conference Series: Earth and Environmental Science*, *272*(3), 32166. https://doi.org/10.1088/1755-1315/272/3/032166

Borning, A., Waddell, P., & Förster, R. (2008). Urbansim: Using Simulation to Inform Public Deliberation and Decision-Making. In R. Sharda, S. Voß, H. Chen, L. Brandt, V. Gregg, R. Traunmüller, S. Dawes, E. Hovy, A. Macintosh, & C. A. Larson (Eds.), *Integrated Series In Information Systems. Digital Government* (Vol. 17, pp. 439–464). Springer US. https://doi.org/10.1007/978-0-387-71611-4_22

Borutta, Y., Haag, M., Hoffmann, H., Kevekordes, J., & Vogt, V. (2020). *Fundamentalkritik des White Papers und des Datenstrategiepapiers der EU-Kommission vom 19. Februar 2020*. https://goal-projekt.de/wp-content/uploads/2020/03/Fundamentalkritik-1.pdf

Botsman, R. (2017). *Who can you trust? How technology brought us together and why it might drive us apart* (First edition (eBook)). PublicAffairs.

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences. Inside technology*. MIT Press.

Boyd, K. (2022). Designing Up with Value-Sensitive Design: Building a Field Guide for Ethical ML Development. In *ICPS, FAccT 2022: 2022 5th ACM Conference on Fairness, Accountability, and Transparency : June 21-24, 2022, Seoul, South Korea* (pp. 2069–2082). The Association for Computing Machinery. https://doi.org/10.1145/3531146.3534626

Bratteteig, T., & Wagner, I. (2012). Disentangling power and decision-making in participatory design. In K. Halskov (Ed.), *ACM Other conferences, Proceedings of the 12th Participatory Design Conference Research Papers - Volume 1* (pp. 41–50). ACM.

Brekke, J. K., & Alsindi, W. Z. (2021). Cryptoeconomics. *Internet Policy Review*, *10*(2). https://doi.org/10.14763/2021.2.1553

Brey, P. (2000). Disclosive computer ethics. *ACM SIGCAS Computers and Society*, *30*(4), 10–16. https://doi.org/10.1145/572260.572264

Brey, P. (2008). The Technological Construction of Social Power. *Social Epistemology*, *22*(1), 71–95. https://doi.org/10.1080/02691720701773551

Brey, P. (2010). Values in Technology and Disclosive Computer Ethics. In L. Floridi (Ed.), *The Cambridge Handbook of Information and Computer Ethics*.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., . . . Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. https://arxiv.org/pdf/1802.07228

Bucher, T. (2013). Objects of intense feeling: The case of the Twitter API. *Computational Culture*(3).

Bundesregierung. (2020). *Technologie soll dem Menschen dienen: KI-Observatorium nimmt Arbeit auf*. https://www.bundesregierung.de/breg-de/aktuelles/ki-oberservatorium-1726794

Buterin, V. (2014). *DAOs, DACs, DAs and More: An Incomplete Terminology Guide*. Ethereum Blog. https://blog.ethereum.org/2014/05/06/daos-dacs-das-and-more-an-incomplete-terminology-guide/

Butts, J. (2017). *Thanks to Misuse, Apps Can't View Mac Addresses on iOS 11*. https://www.macobserver.com/news/product-news/apps-cant-view-mac-addresses-on-ios-11/

Bynum, T. W. (2008). Milestones in the History of Information and Computer Ethics. In K. E. Himma & H. T. Tavani (Eds.), *The Handbook of Information and Computer Ethics* (pp. 25–48). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470281819.ch2

Calabresi, G. (2008). *The Cost of Accidents: A Legal and Economic Analysis*. Yale University Press.

Callon, M. (1984). Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay. *The Sociological Review*, *32*(1_suppl), 196–233. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x

Cavoukian, A. (2011). *Privacy by Design: The 7 Foundational Principles*. Implementation and Mapping of Fair Information Practices. https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf

Cheatham, B., Javanmardian, K., & Samandari, H. (2019). *Confronting the risks of artificial intelligence*. McKinsey Quarterly. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence

Chohan, U. (2018). The Problems of Cryptocurrency Thefts and Exchange Shutdowns. *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.3131702

Christl, W. (2017). *Corporate surveillance in everyday life: How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions*. Cracked Labs.

Christopher, C. M. (2016). The Bridging Model: Exploring the Roles of Trust and Enforcement in Banking, Bitcoin, and the Blockchain. *Nevada Law Journal*, *17*, 139.

Constantinides, P., Henfridsson, O., & Parker, G. G. (2018). Introduction—Platforms and Infrastructures in the Digital Age. *Information Systems Research*, *29*(2), 381–400. https://doi.org/10.1287/isre.2018.0794

Crawford, K., & Schultz, J. (2014). Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms. *Boston College Law Review*, *55*(1), 93–128. https://heinonline.org/HOL/P?h=hein.journals/bclr55&i=93

Crémer, J., Montjoye, Y.-A. d., & Schweitzer, H. (2019). *Competition policy for the digital era*.

Czeskis, A., Dermendjieva, I., Yapit, H., Borning, A., Friedman, B., Gill, B., & Kohno, T. (2010). Parenting from the pocket: value tensions and technical directions for secure and private parent-teen mobile safety. In L. F. Cranor

(Ed.), *Proceedings of the Sixth Symposium on Usable Privacy and Security - SOUPS '10* (p. 1). ACM Press. https://doi.org/10.1145/1837110.1837130

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. In F. Bacchus & C. Sierra (Eds.), *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4691–4697). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2017/654

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

Datenethikkommission. (2019). *Gutachten der Datenethikkommission.* https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf

De Filippi, P. (2017). In Blockchain we Trust: Vertrauenslose Technologie für eine vertrauenslose Gesellschaft. In Rudolf-Augstein-Stiftung (Ed.), *edition suhrkamp: Vol. 2714. Reclaim Autonomy: Selbstermächtigung in der digitalen Weltordnung* (pp. 53–81). Suhrkamp.

De Filippi, P. (2018). *No Blockchain Is an Island.* https://www.coindesk.com/no-blockchain-island/

De Filippi, P., Mannan, M., & Reijers, W. (2020). Blockchain as a confidence machine: The problem of trust & challenges of governance. *Technology in Society*, *62*, 101284. https://doi.org/10.1016/j.techsoc.2020.101284

De Filippi, P., Mannan, M., & Reijers, W. (2022). The alegality of blockchain technology. *Policy and Society*, Article puac006. Advance online publication. https://doi.org/10.1093/polsoc/puac006

De Filippi, P., & Wright, A. (2018). *Blockchain and the law: The rule of code.* Harvard University Press.

Dhinakaran, A. (2020). *The AI Ecosystem is a MESS: Why is it impossible to understand what AI companies really do?* Towards Data Science. https://towardsdatascience.com/the-ai-ecosystem-is-a-mess-c46bdfbf43e4

Digital Europe. (2021). *DIGITALEUROPE's initial findings on the proposed AI Act*. Digital Europe. https://www.digitaleurope.org/wp/wp-content/uploads/2021/08/DIGITALEUROPEs-initial-findings-on-the-proposed-AI-Act.pdf

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.

Dignum, V. (2020). Responsibility and Artificial Intelligence. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *[Oxford handbooks series]. The Oxford handbook of ethics of AI* (pp. 215–231). Oxford University Press.

Dos Santos, R. P. (2017). On the Philosophy of Bitcoin/Blockchain Technology: Is it a Chaotic, Complex System? *Metaphilosophy*, *48*(5), 620–633. https://doi.org/10.1111/meta.12266

Dowding, K. M. (1996). *Power. Concepts in the social sciences series*. Open University Press.

Duhigg, C. (2012). *How Companies Learn Your Secrets*. The New York Times. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

DuPont, Q. (2018). Experiments in Algorithmic Governance: A history and ethnography of "The DAO," a failed Decentralized Autonomous Organization. In M. Campbell-Verduyn (Ed.), *RIPE Series in Global Political Economy. Bitcoin and beyond: Cryptocurrencies, blockchains and global governance*. Routledge.

DuPont, Q. (2019). *Cryptocurrencies and blockchains. Digital media and society series*. Polity.

DuPont, Q., & Maurer, B. (2015). Ledgers and Law in the Blockchain. *Kings Review (23 June 2015) Http://kingsreview. Co. Uk/magazine/blog/2015/06/23/ledgers-and-Law-in-the-Blockchain*.

Eaton, B., Elaluf-Calderwood, S., Sørensen, C., & Yoo, Y. (2015). Distributed Tuning of Boundary Resources: The Case of Apple's iOS Service System. *MIS Quarterly*, *39*(1), 217–243. https://doi.org/10.25300/MISQ/2015/39.1.10

Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H., & Steinrötter, B. (2021). The European Commission’s Proposal for an Artificial Intelligence Act—A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). *J*, *4*(4), 589–603. https://doi.org/10.3390/j4040043

The Economist (Ed.). (2015). *The trust machine*. https://www.economist.com/leaders/2015/10/31/the-trust-machine

European Commission. (2016). *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN

European Commission. (2019). *Building Trust in Human-Centric Artificial Intelligence*. https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

European Commission. (2020a). *On Artificial Intelligence - A European approach to excellence and trust: Whitepaper*. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission. (2020b). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on a Single Market For Digital Services (Digital Services Act) and amending Directive*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en

European Commission. (2020c). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on contestable and fair markets in the* *digital sector (Digital Markets Act)*. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0842&from=en

European Commission. (2020d). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on Markets in Crypto-assets, and amending Directive (EU) 2019/1937*.

European Commission. (2021a). *ANNEXES to the Proposal for a Regulation of the European Parliament and of the Council LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_2&format=PDF

European Commission. (2021b). *COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT Accompanying the Proposal for a Regulation of the European Parliament and of the Council LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. https://eur-lex.europa.eu/resource.html?uri=cellar:0694be88-a373-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

European Commission. (2021c). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

FAT-ML. (2018). *Fairness, Accountability, and Transparency in Machine Learning*. http://www.fatml.org

Feiler, P., Sullivan, K., Wallnau, K., Gabriel, R., Goodenough, J., Linger, R., Longstaff, T., Kazman, R., Klein, M., Northrop, L., & Schmidt, D. (2006). *Ultra-Large-Scale Systems: The Software Challenge of the Future*. Software Engineering Institute, Carnegie Mellon University.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI. *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.3518482

Flanagan, M., Howe, D. C [Daniel C.], & Nissenbaum, H. (2001). Embodying Values in Technology: Theory and Practice. In J. van den Hoven & J. Weckert (Eds.), *Information Technology and Moral Philosophy* (pp. 322–353). Cambridge University Press. https://doi.org/10.1017/CBO9780511498725.017

Floridi, L., & Sanders, J. W. (2001). Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*, *3*(1), 55–66. https://doi.org/10.1023/A:1011440125207

Floridi, L., & Sanders, J. W. (2002). Mapping the foundationalist debate in computer ethics. *Ethics and Information Technology*, *4*(1), 1–9. https://doi.org/10.1023/A:1015209807065

Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Fontana-Giusti, G. (2013). *Foucault for architects*. *Thinkers for architects*. Routledge.

Foucault, M. (2012). *Discipline and Punish: The Birth of the Prison*. Random House US.

Freeman, S., Beveridge, I., & Angelis, J. (2020). Drivers of digital trust in the crypto industry. In M. Ragnedda & G. Destefanis (Eds.), *Routledge studies in science, technology and society. Blockchain and Web 3.0: Social, economic, and technological challenges* (pp. 62–77). Routledge.

Friedman, B., Harbers, M., Hendry, D. G., van den Hoven, J., Jonker, C., & Logler, N. (2021). Eight grand challenges for value sensitive design from the 2016 Lorentz workshop. *Ethics and Information Technology*, *23*(1), 5–16. https://doi.org/10.1007/s10676-021-09586-y

Friedman, B., & Hendry, D. (2019). *Value sensitive design: Shaping technology with moral imagination*. The MIT Press.

Friedman, B., Hendry, D. G., & Borning, A. (2017). A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction*, *11*(2), 63–125. https://doi.org/10.1561/1100000015

Friedman, B., Howe, D. C [D. C.], & Felten, E. (2002). Informed consent in the Mozilla browser: implementing value-sensitive design. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences* (p. 10). IEEE Comput. Soc. https://doi.org/10.1109/HICSS.2002.994366

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value Sensitive Design and Information Systems. In K. E. Himma & H. T. Tavani (Eds.), *The Handbook of Information and Computer Ethics* (pp. 69–101). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470281819.ch4

Friedman, B., Kahn, P., Hagman, J., Severson, R., & Gill, B. (2006). The Watcher and the Watched: Social Judgments About Privacy in a Public Place. *Human-Computer Interaction*, *21*(2), 235–272. https://doi.org/10.1207/s15327051hci2102_3

Friedman, B., Khan, P. H., & Howe, D. C [Daniel C.] (2000). Trust online. *Communications of the ACM*, *43*(12), 34–40. https://doi.org/10.1145/355112.355120

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347. https://doi.org/10.1145/230538.230561

Friedman, B., Smith, I., H. Kahn, P., Consolvo, S., & Selawski, J. (2006). Development of a Privacy Addendum for Open Source Licenses: Value Sensitive Design in Industry. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, P. Dourish, & A. Friday (Eds.), *Lecture Notes in Computer Science. UbiComp 2006: Ubiquitous Computing* (Vol. 4206, pp. 194–211). Springer Berlin Heidelberg. https://doi.org/10.1007/11853565_12

Fu, R., Huang, Y., & Singh, P. V. (2020). Artificial Intelligence and Algorithmic Bias: Source, Detection, Mitigation, and Implications. In C. Druehl, W. Elmaghraby,

D. Shier, & H. J. Greenberg (Eds.), *Pushing the Boundaries: Frontiers in Impactful OR/OM Research* (Vol. 65, pp. 39–63). INFORMS. https://doi.org/10.1287/educ.2020.0215

Gambetta, D. (1988). Can We Trust Trust? In D. Gambetta (Ed.), *Trust: Making and breaking cooperative relations* (pp. 213–237). Blackwell.

Gartenberg, C. (2020). *Spotify, Epic, Tile, Match, and more are rallying developers against Apple's App Store policies: As the 'Coalition for App Fairness'*. https://www.theverge.com/2020/9/24/21453745/spotify-epic-tile-match-coalition-for-app-fairness-apple-app-store-policies-protest

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92.

Ghazawneh, A., & Henfridsson, O. (2013). Balancing platform control and external contribution in third-party development: The boundary resources model. *Information Systems Journal*, *23*(2), 173–192. https://doi.org/10.1111/j.1365-2575.2012.00406.x

Giungato, P., Rana, R., Tarabella, A., & Tricase, C. (2017). Current Trends in Sustainability of Bitcoins and Related Blockchain Technology. *Sustainability*, *9*(12), 2214. https://doi.org/10.3390/su9122214

Google. (2021). *Consultation on the EU AI Act Proposal: Google's submission*. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2662492_en

Gotterbarn, D. (1991). Computer ethics: responsibility regained. *National Forum: The Phi Beta Kappa Journal*, 26–31.

Granka, L. A. (2010). The Politics of Search: A Decade Retrospective. *The Information Society*, *26*(5), 364–374. https://doi.org/10.1080/01972243.2010.511560

Greene, D., & Shilton, K. (2017). Platform privacies: Governance, collaboration, and the different meanings of "privacy" in iOS and Android development. *New Media & Society*, *126*(3), 1-18. https://doi.org/10.1177/1461444817702397

Greenfield, A. (2017). *Radical technologies: The design of everyday life*. Verso.

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, *7*, 47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068

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

Hansen, H. R., Mendling, J., & Neumann, G. (2019). *Wirtschaftsinformatik: Grundlagen und Anwendungen* (12. völlig neu bearbeitete Auflage). *De Gruyter Oldenbourg : Studium*. De Gruyter.

Hao, K. (2021). *The Facebook whistleblower says its algorithms are dangerous. Here's why*. MIT Technology Review. https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/

Hardin, R. (2002). *Trust and trustworthiness. The Russell Sage Foundation series on trust: volume 4*. Russell Sage Foundation.

Haugaard, M. (2010). Power: A 'family resemblance' concept. *European Journal of Cultural Studies*, *13*(4), 419–438. https://doi.org/10.1177/1367549410377152

Hawley, K. (2017). Trustworthy Groups and Organizations. In P. Faulkner & T. Simpson (Eds.), *The Philosophy of Trust* (pp. 230–250). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198732549.003.0014

Hawlitschek, F., Notheisen, B., & Teubner, T. (2018). The limits of trust-free systems: A literature review on blockchain technology and trust in the sharing economy. *Electronic Commerce Research and Applications*, *29*, 50–63. https://doi.org/10.1016/j.elerap.2018.03.005

Hein, A., Schreieck, M., Riasanow, T., Setzke, D. S., Wiesche, M., Böhm, M., & Krcmar, H. (2020). Digital platform ecosystems. *Electronic Markets*, *30*(1), 87–98. https://doi.org/10.1007/s12525-019-00377-4

Hestres, L. (2013). App neutrality: Apple's app store and freedom of expression online. *Hestres, LE (2013). App Neutrality: Apple's App Store and Freedom of Expression Online. International Journal of Communication*, *7*, 1265–1280.

Hill, K. (2012). *How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did*. Forbes. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/?sh=20f35caa6668

HLEG-AI. (2018). *A Definition of AI: Main Capabilities and Scientific Disciplines*.

HLEG-AI. (2019). *Ethics guidelines for trustworthy AI*.

Hoffman, R. (2014). *The Future of the Bitcoin Ecosystemd and "Trustless Trust": Why I Invested in Blockstream*. https://www.linkedin.com/pulse/20141117154558-1213-the-future-of-the-bitcoin-ecosystem-and-trustless-trust-why-i-invested-in-blockstream

Hoffmann, H. (2021). Regulierung der Künstlichen Intelligenz: Fundamentalkritik am Verordnungsentwurf zur Regulierung der Künstlichen Intelligenz der EU-Kommission vom 21. 4. 2021. *Kommunikation & Recht*, 369–374.

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy*, *72*(1), 63–76. https://doi.org/10.1080/00048409412345881

Horowitz, M. C. (2016). The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons. *Daedalus*, *145*(4), 25–36. https://doi.org/10.1162/DAED_a_00409

Introna, L. D. (2005). Disclosive Ethics and Information Technology: Disclosing Facial Recognition Systems. *Ethics and Information Technology*, *7*(2), 75–86. https://doi.org/10.1007/s10676-005-4583-2

Jacobs, M. (2021). How Implicit Assumptions on the Nature of Trust Shape the Understanding of the Blockchain Technology. *Philosophy & Technology*, *34*(3), 573–587. https://doi.org/10.1007/s13347-020-00410-x

Jacobs, M., Kurtz, C., Simon, J., & Böhmann, T. (2021). Value Sensitive Design and power in socio-technical ecosystems. *Internet Policy Review*, *10*(3). https://doi.org/10.14763/2021.3.1580

Jacobs, M., & Simon, J. (2022a). Assigning Obligations in AI Regulation: A Discussion of Two Frameworks Proposed by the European Commission. *Digital Society*, *1*(1), Article 6. https://doi.org/10.1007/s44206-022-00009-z

Jacobs, M., & Simon, J. (2022b). Reexamining computer ethics in light of AI systems and AI regulation. *AI and Ethics*. Advance online publication. https://doi.org/10.1007/s43681-022-00229-6

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

Johnson, D. G. (1999). Sorting out the uniqueness of computer-ethical issues.

Johnson, D. G., & Miller, K. W. (2009). *Computer ethics: Analyzing information technology* (4. ed.). Pearson Education Intern.

Jones, K. (1996). Trust as an Affective Attitude. *Ethics*, *107*(1), 4–25. https://doi.org/10.1086/233694

Jonsson, K. (2006). The embedded panopticon: visibility issues of remote diagnostics surveillance. *Scandinavian Journal of Information Systems*, *18*(2), 3.

Karhu, K., Gustafsson, R., & Lyytinen, K. (2018). Exploiting and defending open digital platforms with boundary resources: Android's five platform forks. *Information Systems Research : ISR : An Information Systems Journal of the Institute for Operations Research and the Management Sciences*, *29*(2), 479–497.

Keller, J. R., Chauvet, L., Fawcett, J., & Thereaux, O. (2018). *The role of data in AI business models*. Open Data Institute. https://theodi.org/wp-content/uploads/2018/04/376886336-The-role-of-data-in-AI-business-models.pdf

Kemppainen, L., Pikkarainen, M., Hurmelinna-Laukkanen, P., & Reponen, J. (2019). Data Access in Connected Health Innovation: Managerial Orchestration Challenges and Solutions. *Technology Innovation Management Review*, *9*(12), 43–55. https://doi.org/10.22215/timreview/1291

Keyes, O., Hutson, J., & Durbin, M. (2019). A Mulching Proposal. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–11). ACM. https://doi.org/10.1145/3290607.3310433

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores*. https://arxiv.org/pdf/1609.05807

Kommission Wettbewerbsrecht 4.0. (2019). *A New Competition Framework for the Digital Economy: Report by the Commission 'Competition Law 4.0'*. https://www.bmwi.de/Redaktion/EN/Downloads/a/a-new-competition-framework.pdf?__blob=publicationFile&v=2

Krafft, T. D., Zweig, K. A., & König, P. D. (2020). How to regulate algorithmic decision-making: A framework of regulatory requirements for different applications. *Regulation & Governance*, *104*, 119–136. https://doi.org/10.1111/rego.12369

KRAKENFX. (2019). *Kraken is Delisting BSV*. https://blog.kraken.com/post/2274/kraken-is-delisting-bsv/

Krishna, A. (2015). *Blockchain: It Really is a Big Deal*. https://www.ibm.com/blogs/think/2015/09/blockchain-really-big-deal/

Kroes, P., Franssen, M., van Poel, I. de, & Ottens, M. (2006). Treating socio-technical systems as engineering systems: some conceptual problems. *Behavioral Science*, *23*(6), 803–814. https://doi.org/10.1002/sres.703

Kurtz, C., Semmann, M., & Schulz, W. (2018). Towards a framework for information privacy in complex service ecosystems.

Kurtz, C., Wittner, F., Semmann, M., Schulz, W., & Böhmann, T. (2022). Accountability of platform providers for unlawful personal data processing in their ecosystems–A socio-techno-legal analysis of Facebook and Apple's iOS according to GDPR. *Journal of Responsible Technology*, *9*. https://doi.org/10.1016/j.jrt.2021.100018

Latour, B. (1992). Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts. In W. E. Bijker & J. Law (Eds.), *Inside technology. Shaping technology/building society: Studies in sociotechnical change* (pp. 225–258). MIT Press.

Law, J., & Callon, M. (1992). The Life and Death of an Aircraft: A Network Analysis of Technical Change. In W. E. Bijker & J. Law (Eds.), *Inside technology. Shaping* *technology/building society: Studies in sociotechnical change* (pp. 21–52). MIT Press.

Leerssen, P., Ausloos, J., Zarouali, B., Helberger, N., & Vreese, C. H. de (2019). Platform ad archives: promises and pitfalls. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1421

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

Leswing, K. (2021). *Apple exec warns it may remove apps that track users without permission.* CNBC. https://www.cnbc.com/2020/12/08/apple-may-remove-apps-that-track-users-without-permission-in-2021.html

Li, T., Louie, E., Dabbish, L., & Hong, J. I. (2021). How Developers Talk About Personal Data and What It Means for User Privacy. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1–28. https://doi.org/10.1145/3432919

Locke, J. (1988). Two treatises of government, ed. *Peter Laslett (Cambridge, 1988), 301.*

Lowry, S., & Macpherson, G. (1988). A blot on the profession. *British Medical Journal (Clinical Research Ed.)*, *296*(6623), 657–658. https://doi.org/10.1136/bmj.296.6623.657

Lusch, R. F., & Nambisan, S. (2015). Service Innovation: A Service-Dominant Logic Perspective. *MIS Quarterly*, *39*(1), 155–175. https://doi.org/10.25300/MISQ/2015/39.1.07

Mallard, A., Méadel, C., & Musiani, F. (2014). The paradoxes of distributed trust: Peer-to-peer architecture and user confidence in Bitcoin. *Journal of Peer Production*(4), 1–10.

Maner, W. (1999). Is computer ethics unique?

Maurer, B., Nelms, T. C., & Swartz, L. (2013). "When perhaps the real problem is money itself! ": the practical materiality of Bitcoin. *Social Semiotics*, *23*(2), 261–277. https://doi.org/10.1080/10350330.2013.777594

McConahy, A., Eisenbraun, B., Howison, J., Herbsleb, J. D., & Sliz, P. (2012). Techniques for monitoring runtime architectures of socio-technical ecosystems. In *Workshop on Data-Intensive Collaboration in Science and Engineering (CSCW 2012)*.

McLeod, C. (2006). *Trust*. https://plato.stanford.edu/entries/trust

Microsoft. (2018). *Pre-trained machine learning models for sentiment analysis and image detection*. Microsoft. https://docs.microsoft.com/en-us/machine-learning-server/install/microsoftml-install-pretrained-models

Milano, S., Taddeo, M., & Floridi, L. (2020). Ethical aspects of multi-stakeholder recommendation systems. *The Information Society*, 1–11. https://doi.org/10.1080/01972243.2020.1832636

Miller, J. K., Friedman, B., & Jancke, G. (2007). Value tensions in design. In T. Gross & K. Inkpen (Eds.), *Proceedings of the 2007 international ACM conference on Conference on supporting group work - GROUP '07* (p. 281). ACM Press. https://doi.org/10.1145/1316624.1316668

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

Monopolkommission. (2015). *Competition policy: The challenge of digital markets*. Special Report by the Monopolies Commission pursuant to section 44(1)(4) of the Act Against Restraints on Competition. https://www.monopolkommission.de/images/PDF/SG/s68_fulltext_eng.pdf

Moor, J. H. (1985). What Is Computer Ethics? *Metaphilosophy*, *16*(4), 266–275. https://doi.org/10.1111/j.1467-9973.1985.tb00173.x

Moor, J. H. (2001). The future of computer ethics: You ain't seen nothin' yet! *Ethics and Information Technology*, *3*(2), 89–91. https://doi.org/10.1023/A:1011881522593

Moor, J. H. (2005). Why We Need Better Ethics for Emerging Technologies. *Ethics and Information Technology*, *7*(3), 111–119. https://doi.org/10.1007/s10676-006-0008-0

Morabito, V. (2017). *Business Innovation Through Blockchain: The B³ Perspective*. Springer International Publishing.

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine (1982)*, *260*, 113172. https://doi.org/10.1016/j.socscimed.2020.113172

Mueller, M., & Heger, O. (2018). Health at any Cost? Investigating Ethical Dimensions and Potential Conflicts of an Ambulatory Therapeutic Assistance System through Value Sensitive Design. In *39th International Conference on Information Systems (ICIS).*

Müller, V. C. (2020). *Ethics of Artificial Intelligence and Robotics*. https://plato.stanford.edu/entries/ethics-ai/

Munn, L. (2022). The uselessness of AI ethics. *AI and Ethics*. Advance online publication. https://doi.org/10.1007/s43681-022-00209-w

Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*. https://bitcoin.org/bitcoin.pdf

Nathan, L. P. (2012). Sustainable information practice: An ethnographic investigation. *Journal of the American Society for Information Science and Technology*, *63*(11), 2254–2268. https://doi.org/10.1002/asi.22726

Nathan, L. P., Klasnja, P. V., & Friedman, B. (2007). Value scenarios. In M. B. Rosson & D. Gilmore (Eds.), *CHI '07 extended abstracts on Human factors in computing systems - CHI '07* (p. 2585). ACM Press. https://doi.org/10.1145/1240866.1241046

Neff, G., & Nagy, P. (2016). Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication*.

Nickel, P. J. (2013). Trust in Technological Systems. In M. J. Vries, S. O. Hansson, & A. W. Meijers (Eds.), *Philosophy of Engineering and Technology: Vol. 9. Norms in Technology* (pp. 223–237). Springer.

Nieborg, D., Poell, T., & van Dijck, J. (2020). Analyzing Platform Power: App Stores as Infrastructural Platform Services. In *Selected Papers of #AoIR2019: The*

20th Annual Conference of the Association of Internet Researchers. https://doi.org/10.5210/spir.v2019i0.11019

Nissenbaum, H. (1994). Computing and accountability. *Communications of the ACM*, *37*(1), 72–80. https://doi.org/10.1145/175222.175228

Nissenbaum, H. (2005). Values in Technical Design. In C. Mitcham (Ed.), *Encyclopedia of science, technology, and ethics* (pp. lxvi–lxx). Macmillan Reference USA.

Nozick, R. (2001). *Anarchy, State, and Utopia*. Wiley-Blackwell.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First paperback edition). Broadway Books.

Orcutt, M. (2018). *Chain Letter #102: You can go your own way*. https://mailchi.mp/technologyreview/chain-letter-767541?e=93dc606e34&utm_campaign=chain_letter.unpaid.engagement&utm_source=hs_email&utm_medium=email&utm_content=71892724&_hsenc=p2ANqtz--2S3Tqvwcm1BQc_Eb-ArlTKDA-f9cAdBdc6Lxq67nmBy3Y2Z48OyMOEMW__mczpT4YRw2w-7sUtAvryWgLnYMUo_VkOxYs6DqYzXyno1pRK7AGncM&_hsmi=71892724

Orcutt, M. (2019). *Chain Letter #139: On social media-fueled coin delistings*. https://cryptocurrency-news-now.blogspot.com/2019/04/139-on-social-media-fueled-coin.html

Perez, S. (2020). *Coalition for App Fairness, a group fighting for app store reforms, adds 20 new partners*. https://techcrunch.com/2020/10/21/coalition-for-app-fairness-a-group-fighting-for-app-store-reforms-adds-20-new-partners/

Poell, T., Nieborg, D., & van Dijck, J. (2019). Platformisation. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1425

Quan, X. I., & Sanderson, J. (2018). Understanding the Artificial Intelligence Business Ecosystem. *IEEE Engineering Management Review*, *46*(4), 22–25. https://doi.org/10.1109/EMR.2018.2882430

Rawls, J. (1999). *A Theory of Justice: Revised Edition* (Rev Sub). Belknap Press.

Reijers, W., O'Brolcháin, F., & Haynes, P. (2016). Governance in Blockchain Technologies & Social Contract Theories. *Ledger*, *1*, 134–151. https://doi.org/10.5195/ledger.2016.62

Reuver, M. de, Sørensen, C., & Basole, R. C. (2018). The Digital Platform: A Research Agenda. *Journal of Information Technology*, *33*(2), 124–135. https://doi.org/10.1057/s41265-016-0033-3

Rieder, G., Simon, J., & Wong, P.-H. (2021). Mapping the Stony Road toward Trustworthy AI: Expectations, Problems, Conundrums. In M. Pelillo & T. Scantamburlo (Eds.), *Machines we trust: Perspectives on dependable AI*. The MIT Press. https://doi.org/10.7551/mitpress/12186.003.0007

Ropohl, G. (1999). Philosophy of Socio-Technical Systems. *Society for Philosophy and Technology Quarterly Electronic Journal*, *4*(3), 186–194. https://doi.org/10.5840/techne19994311

Ross, C., & Swetlitz, I. (2018). *IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show*. https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf

Rousseau, J.-J., & Cole, G. D. H. (1992). *The social contract and discourses* (Repr [der Ausg.] London 1973). *Everyman's library*. Dent.

Rudschies, C. (2023). Exploring the Concept of Solidarity in the Context of AI: An Ethics in Design Approach. *Digital Society*, *2*(1). https://doi.org/10.1007/s44206-022-00027-x

Russell, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. *Prentice Hall series in artificial intelligence*. Prentice Hall; London : Prentice-Hall International.

Sartori, L., & Theodorou, A. (2022). A sociotechnical perspective for the future of AI: narratives, inequalities, and human control. *Ethics and Information Technology*, *24*(1). https://doi.org/10.1007/s10676-022-09624-3

Sattarov, F. (2019). *Power and technology: A philosophical and ethical analysis*. Rowman et Littlefield.

Saulles, M. de, & Horner, D. S. (2011). The portable panopticon: morality and mobile technologies. *Journal of Information, Communication and Ethics in Society*, *9*(3), 206–216. https://doi.org/10.1108/14779961111167676

Sclove, R. E. (1992). The Nuts and Bolts of Democracy: Democratic Theory and Technological Design. In L. Winner (Ed.), *Democracy in a Technological Society* (pp. 139–157). Springer Netherlands. https://doi.org/10.1007/978-94-017-1219-4_9

Shilton, K. (2012). Values Levers. *Science, Technology, & Human Values*, *38*(3), 374–397. https://doi.org/10.1177/0162243912436985

Shilton, K., & Greene, D. (2016). Because Privacy: Defining and Legitimating Privacy in iOS Development. In X. Lin & M. Khoo (Eds.), *iConference 2016 Proceedings*. iSchools. https://doi.org/10.9776/16229

Shilton, K., & Greene, D. (2019). Linking Platforms, Practices, and Developer Ethics: Levers for Privacy Discourse in Mobile Application Development. *Journal of Business Ethics*, *155*(1), 131–146. https://doi.org/10.1007/s10551-017-3504-8

Simon, J. (2013). Trust. In D. Pritchard (Ed.), *Oxford Bibliographies*. Oxford University Press. https://doi.org/10.1093/obo/9780195396577-0157

Simon, J. (2016). Values in Design. In J. Heesen (Ed.), *Handbuch Medien- und Informationsethik* (pp. 357–364). J.B. Metzler.

Simon, J., Wong, P.-H., & Rieder, G. (2020). Algorithmic bias and the Value Sensitive Design approach. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1534

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701–717. https://doi.org/10.1006/ijhc.1999.0349

Slota, S. C. (2020). Designing Across Distributed Agency: Values, participatory design and building socially responsible AI. *Good Systems-Published Research*.

Smith, T. D. (2017). The blockchain litmus test. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 2299–2308). IEEE. https://doi.org/10.1109/BigData.2017.8258183

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act. *SSRN Electronic Journal*. Advance online publication. https://doi.org/10.2139/ssrn.3899991

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Stahl, B. C. (2022). From computer ethics and the ethics of AI towards an ethics of digital ecosystems. *AI and Ethics*, *2*(1), 65–77. https://doi.org/10.1007/s43681-021-00080-1

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, `Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*(3), 387–420. https://doi.org/10.1177/030631289019003001

Strahilevitz, L. J. (2008). Privacy versus antidiscrimination. *The University of Chicago Law Review*, *75*(1), 363–381.

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, *56*(4), 24–31. https://doi.org/10.1109/MSPEC.2019.8678513

Swan, M. (2015). *Blockchain: Blueprint for a New Economy*. O'Reilly.

Swan, M., & De Filippi, P. (2017). Toward a Philosophy of Blockchain: A Symposium: Introduction. *Metaphilosophy*, *48*(5), 603–619. https://doi.org/10.1111/meta.12270

Szabo, N. (2017). *Money, blockchains, and social scalability*. https://unenumerated.blogspot.com/2017/02/money-blockchains-and-social-scalability.html

Tauro, C. K. (2021). Values of privacy and trust for monitoring health in injecting drug users. In *2021 IEEE 25th International Enterprise Distributed Object*

Computing Workshop (EDOCW) (pp. 95–102). IEEE. https://doi.org/10.1109/EDOCW52865.2021.00038

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Tiwana, A., & Konsynski, B. (2010). Complementarities Between Organizational IT Architecture and Governance Structure. *Information Systems Research*, *21*(2), 288–304. https://doi.org/10.1287/isre.1080.0206

Tiwana, A., Konsynski, B., & Bush, A. A. (2010). Research Commentary —Platform Evolution: Coevolution of Platform Architecture, Governance, and Environmental Dynamics. *Information Systems Research*, *21*(4), 675–687. https://doi.org/10.1287/isre.1100.0323

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: key problems and solutions. *AI & SOCIETY*, *37*(1), 215–230. https://doi.org/10.1007/s00146-021-01154-8

Umbrello, S. (2019). Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach. *Big Data and Cognitive Computing*, *3*(1), 5. https://doi.org/10.3390/bdcc3010005

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, *1*(3), 283–296. https://doi.org/10.1007/s43681-021-00038-3

Underwood, S. (2016). Blockchain beyond bitcoin. *Communications of the ACM*, *59*(11), 15–17. https://doi.org/10.1145/2994581

Vallor, S., & Bekey, G. A. (2017). Artificial Intelligence and the Ethics of Self-learning Robots. In P. Lin, R. Jenkins, & K. Abney (Eds.), *Robot ethics 2.0: New challenges in philosophy, law, and society* (pp. 338–353). Oxford University Press.

van Alstyne, M., Parker, G., & Choudary, S. P. (2016). Pipelines, platforms, and the new rules of strategy. *Harvard Business Review : HBR*, *94*(4), 54–62.

van den Hoven, J. (2008). Moral Methodology and Information Technology. In K. E. Himma & H. T. Tavani (Eds.), *The Handbook of Information and Computer*

Ethics (pp. 49–67). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470281819.ch3

van Dijck, J., Poell, T., & Waal, M. de. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

van House, N. A. (2004). Science and technology studies and information studies. *Annual Review of Information Science and Technology (ARIST)*, *38*, 3–86.

van Maanen, G. (2022). Ai Ethics, Ethics Washing, and the Need to Politicize Data Ethics. *Digital Society : Ethics, Socio-Legal and Governance of Digital Technology*, *1*(2), 9. https://doi.org/10.1007/s44206-022-00013-3

van Rest, J., Boonstra, D., Everts, M., van Rijn, M., & van Paassen, R. (2014). Designing Privacy-by-Design. In B. Preneel & D. Ikonomou (Eds.), *Lecture Notes in Computer Science: Vol. 8319. Privacy Technologies and Policy: First Annual Privacy Forum, APF 2012, Limassol, Cyprus, October 10-11, 2012, Revised Selected Papers* (Vol. 8319, pp. 55–72). Springer Berlin Heidelberg; Imprint; Springer. https://doi.org/10.1007/978-3-642-54069-1_4

Veale, M., & Zuiderveen Borgesius, F. (2021). *Demystifying the Draft EU Artificial Intelligence Act*. Pre-print, July 2021. Version 1.1. https://doi.org/10.31235/osf.io/38p5f

Velasco, P. R. (2017). Computing Ledgers and the Political Ontology of the Blockchain. *Metaphilosophy*, *48*(5), 712–726. https://doi.org/10.1111/meta.12274

Vermaas, P. E., Tan, Y., van den Hoven, J., Burgemeestre, B., & Hulstijn, J. (2010). Designing for Trust: A Case of Value-Sensitive Design. *Knowledge, Technology & Policy*, *23*(3-4), 491–505. https://doi.org/10.1007/s12130-010-9130-8

Voss, T., & Tutic, A. (2020). Trust and Game Theory. In J. Simon (Ed.), *The Routledge Handbook of Trust an Philosophy*. Routledge.

Waddell, P., Wang, L., & Liu, X. (2008). UrbanSim: An evolving planning support system for evolving communities. *Planning Support Systems for Cities and Regions*, 103–138.

Wagner, B. (2019). Ethics As An Escape From Regulation. From “Ethics-Washing” To Ethics-Shopping? In E. Bayamlioglu, I. Baraliuc, L. A. W. Janssens, & M.

Hildebrandt (Eds.), *BEING PROFILED: COGITAS ERGO SUM: 10 Years of Profiling the European Citizen* (pp. 84–89). Amsterdam University Press. https://doi.org/10.1515/9789048550180-016

Walch, A. (2019a). Deconstructing 'Decentralization': Exploring the Core Claim of Crypto Systems. *SSRN Electronic Journal*.

Walch, A. (2019b). In Code(rs) We Trust: Software Developers as Fiduciaries in Public Blockchains. In I. Lianos, P. Hacker, S. Eich, & G. Dimitropoulos (Eds.), *Regulating Blockchain: Techno-Social and Legal Challanges* (pp. 58–81). Oxford University Press.

Walton, R., & DeRenzi, B. (2009). Value-Sensitive Design and Health Care in Africa. *IEEE Transactions on Professional Communication*, *52*(4), 346–358. https://doi.org/10.1109/TPC.2009.2034075

Warnier, M., Dechesne, F., & Brazier, F. (2015). Design for the Value of Privacy. In J. van den Hoven, Vermaas Pieter E., & I. van de Poel (Eds.), *Springer reference. Handbook of ethics, values, and technological design: Sources, theory, values and application domains* (pp. 431–445). Springer.

Weber, M. (2019). *Economy and Society: A New Translation*. Harvard University Press.

Webster, K., Wang, X., Tenney, I., Beutel, A., Pitler, E., Pavlick, E., Chen, J., Chi, E., & Petrov, S. (2020). *Measuring and Reducing Gendered Correlations in Pre-trained Models*. http://arxiv.org/pdf/2010.06032v2

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. Freeman.

Werbach, K. (2017). Trust, But Verify: Why the Blockchain Needs the Law. *Berkeley Technology Law Journal*.

Werbach, K. (2018). *The blockchain and the new architecture of trust*. *Information policy series*. The MIT Press.

Wiener, N. (1961). *Cybernetics or control and communication in the animal and the machine* (2. ed., 14. print). MIT Press.

Wiener, N. (1989). *The human use of human beings: Cybernetics and society*. Free Association.

Willemink, M. J., Koszek, W. A., Hardell, C., Wu, J., Fleischmann, D., Harvey, H., Folio, L. R., Summers, R. M., Rubin, D. L., & Lungren, M. P. (2020). Preparing Medical Imaging Data for Machine Learning. *Radiology*, *295*(1), 4–15. https://doi.org/10.1148/radiol.2020192224

Winkler, T., & Spiekermann, S. (2018). Twenty years of value sensitive design: A review of methodological practices in VSD projects. *Ethics and Information Technology*, *18*(4), 185. https://doi.org/10.1007/s10676-018-9476-2

Winner, L. (1980). Do artifacts have politics? *Daedalus*, 121–136.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*.

Wolf, M. J., Miller, K. W., & Grodzinsky, F. S. (2017). Why We Should Have Seen That Coming: Comments on Microsoft's Tay "Experiment," and Wider Implications. *The ORBIT Journal*, *1*(2), 1–12. https://doi.org/10.29297/orbit.v1i2.49

Wright, A., & De Filippi, P. (2015). Decentralized blockchain technology and the rise of lex cryptographia. *SSRN Electronic Journal*.

Xu, X., Weber, I., & Staples, M. (2019). *Architecture for Blockchain Applications*. Springer International Publishing; Imprint: Springer.

Yeung, K. (2019). Why Worry about Decision-Making by Machine? In K. Yeung & M. Lodge (Eds.), *Algorithmic regulation* (pp. 21–48).

Yoo, D., Derthick, K., Ghassemian, S., Hakizimana, J., Gill, B., & Friedman, B. (2016). Multi-lifespan Design Thinking. In J. Kaye, A. Druin, C. Lampe, D. Morris, & J. P. Hourcade (Eds.), *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 4423–4434). ACM. https://doi.org/10.1145/2858036.2858366

Zimmerman, A. D. (1995). Toward a more democratic ethic of technological governance. *Science, Technology, & Human Values*, *20*(1), 86–107.

Zweig, K. A., Wenzelburger, G., & Krafft, T. D. (2018). On Chances and Risks of Security Related Algorithmic Decision Making Systems. *European Journal for Security Research*, *3*(2), 181–203. https://doi.org/10.1007/s41125-018-0031-2