National and Kapodistrian University of Athens

Department of History & Philosophy of Science

Explainable Artificial Intelligence: An STS perspective

THESIS

ORFEAS MENIS - MASTROMICHALAKIS

Supervisor : Stathis Psillos
Professor

Athens, August 2023

National and Kapodistrian University of Athens
Department of History & Philosophy of Science

Explainable Artificial Intelligence: An STS perspective

THESIS

ORFEAS MENIS - MASTROMICHALAKIS

Supervisor : Stathis Psillos
Professor

Approved by the committee on August 31, 2023.

....	....	....
Stathis Psillos Professor	Aristotle Tympas Professor	Manolis Simos Ph.D.

Athens, August 2023

....

Orfeas Menis - Mastromichalakis

The views and conclusions contained in this document express the author and should not be interpreted as representing the official positions of the National and Kapodistrian University of Athens.

Abstract

This thesis undertakes a comprehensive exploration of Explainable Artificial Intelligence (XAI) by synergizing perspectives from both technical AI research and Science, Technology, and Society - Science and Technology Studies (STS). Anchored in a thorough literature review, we dissect multifaceted narratives emerging from STS literature on XAI, emphasizing the societal, ethical, and philosophical dimensions of explainability. We then navigate technical avenues, diving into the methods and critiques emanating from the AI community itself. At the core of our discourse is the understanding that AI systems are not merely technical entities but intrinsically woven into the fabric of societal structures and politics. We address the pressing need to perceive AI in terms not only of algorithmic transparency but also of its alignment with human cognition, societal norms, and power dynamics. Through this lens, we unravel the challenges of unclear terminology, the absence of universal objectives, and the intricate interplay between transparency, trust, and interpretability in XAI. Real-world case studies, spanning from music recommendation systems to AI in oncology, offer tangible illustrations, bridging theoretical insights with practical scenarios. These narratives act as crucibles to test and validate our interdisciplinary approach, emphasizing the significance of user-centric designs and the politics embedded within AI systems. By synthesizing these analyses, we illuminate a path towards a more integrated, holistic, and informed approach to XAI—one that champions both technical rigor and societal resonance. As we find ourselves in an era where AI continues to reshape our world, this work sets the stage for more responsible, nuanced, and inclusive advancements in the realm of explainability.

Key words

XAI, STS, AI Transparency, Explainability, Interpretability

Acknowledgements

First and foremost, I would like to thank my supervising committee. I thank Prof. Stathis Psillos for doing me the honor of overseeing this thesis and entrusting me with this academic endeavor. However, along with Stathis, Prof. Telly Tympas, and Dr. Manolis Simos have been the foundational pillars of this work. Their guidance has been invaluable at every turn of this journey. Telly has been a true mentor throughout my time in the STS Master's program, providing unwavering support from the onset. Manolis has been instrumental, offering both inspiration and insightful discussions that enriched my perspectives.

I'm truly thankful to my parents and brother for their support. Their encouragement has been key throughout this journey. My brother, in particular, has been a subtle inspiration, motivating me to follow my aspirations with determination.

I'm grateful to my friends for always being there and providing understanding when needed most. Their support has meant a lot to me. Additionally, a big thanks to D. for lighting up my days and standing by me through any challenge that came our way.

Orfeas Menis - Mastromichalakis,

Athens, August 31, 2023

Contents

List of Figures

Chapter 1

Introduction

Artificial Intelligence (AI) has become an integral force in contemporary society, as it influences and revolutionizes numerous sectors of our daily lives in unprecedented ways. It's not just about machines performing tasks; it's about these systems making crucial decisions, sometimes even without human intervention. As AI systems permeate various fields, from healthcare and finance to areas like autonomous driving, their inherent complexity and decision-making capabilities bring forth a myriad of challenges. Notably, there's a rising concern: How can we ensure these AI-driven decisions are transparent and accountable? Especially in critical domains, the stakes are high, and understanding the rationale behind an AI's decision could be pivotal. In response to these challenges, the research domain of eXplainable Artificial Intelligence (XAI) has emerged. XAI endeavors to demystify the often opaque nature of AI algorithms, ensuring they are not just potent but also transparent and trustworthy. Through XAI, researchers and practitioners alike aspire to bridge the comprehension gap, ensuring AI systems are both effective and understood by those they serve.

The realm of explainable AI has been, for the most part, dominated by the development and refinement of technical methods in recent years. Researchers have endeavored to devise algorithms that can furnish post-hoc clarifications for the often elusive outputs of AI systems, or provide technical solutions to the transparency of AI systems. Yet, as with any burgeoning technology, merely understanding the mechanics isn't enough. The real-world implications and acceptability of such systems hinge on how they resonate with the fabric of society. It is in this intricate blend of the social, cultural, and ethical domains that the true essence of AI's explainability emerges. Recognizing this, our thesis adopts a distinctive lens, one that shifts focus from the purely technical to the societal. By leveraging insights from Science, Technology, and Society - Science and Technology Studies (STS), we aim to dissect the nuances of explainable AI, placing emphasis on its social dimensions and the context in which it operates. This approach promises a more holistic and enriched understanding of AI explainability, going beyond algorithms to the very heart of human-AI interactions.

STS operates at the intersection of science, technology, and society, offering a rich tapestry of insights into their complex interplay. As an interdisciplinary field, STS doesn't merely observe technology in isolation but sees it as deeply rooted in, and reflective of, the societal and cultural contexts from which it emerges. The lens provided by STS encourages a comprehensive evaluation of technological artifacts, illuminating the nuanced interactions they share with prevailing social practices. This emphasis on situating technologies within broader sociotechnical systems provides a depth of understanding often overlooked in more technocentric studies. When we apply the principles of STS to the domain of explainable AI, a transformative shift in perspective occurs. No longer do we just see algorithms and models; instead, we delve into the intricate socio-cultural tapestry that underpins AI's development and application. Through this lens, we can scrutinize how AI systems are not merely technical tools but also socially constructed entities, inherently influenced by human values. Furthermore, we gain the means to discern how these systems reciprocally shape societal norms and expectations. Ultimately, by harnessing the insights of STS, our exploration of AI explainability becomes richer, considering the multifaceted implications for stakeholders ranging from developers to end-users and policymakers.

In the rapidly evolving landscape of Artificial Intelligence, the intricacies of explainability have emerged as both a challenge and a necessity. This thesis embarks on a journey to critically dissect the concept of explainable AI but with a twist. Instead of the oft-trodden technical pathways, our exploration is guided by the nuanced perspectives of STS. By centering our analysis on the social and ethical facets of AI explainability, we move beyond mere algorithmic details and venture into the realm of human values, societal norms, and cultural contexts. The intention is to illuminate the multifaceted benefits, as well as the inherent challenges, of integrating STS viewpoints into the lifecycle of explainable AI systems—from their inception and design to their deployment and real-world applications. As we traverse this path, it's evident that AI is not an isolated technological marvel—it's deeply interwoven with the fabric of society. Through our research, we aspire to enrich the burgeoning literature on AI's societal intersections. Emphasizing the indelible link between AI and its sociotechnical environment, we advocate for a more holistic approach to AI research and development, one that recognizes and respects the profound impacts of these systems on society at large.

To achieve these objectives, this thesis will follow a structured approach, beginning with a comprehensive literature review on STS approaches to explainable AI. Then, we will delve deeper into the technical approaches of XAI, reviewing related literature and discussing critique that emerges within the technical community. This review serves as a foundational bedrock, tracing the historical evolution, key debates, and significant advancements in both domains. By intertwining the insights from AI research with the broader narratives from STS, we aim to identify gaps, synergies, and potential avenues for interdisciplinary inquiry. After the literature review, we combine the findings and discuss the insights gained from it by utilizing useful case studies and empirical analyses when necessary in order to delve deeper into real-world applications and implications of explainable AI. These case studies not only provide tangible examples but also serve as a bridge to connect theoretical findings with practical scenarios. Throughout this process, continuous reflection and engagement with the STS framework will guide our understanding and interpretation. By the culmination of this thesis, we aim to have provided a cohesive, enriched perspective on explainable AI, one that resonates with both technical and societal considerations, thereby paving the way for more responsible and informed AI developments in the future.

To attain the outlined objectives, this thesis embarks on a systematic exploration, starting with an exhaustive literature review of STS perspectives on explainable AI. This is followed by a deep dive into the technical dimensions of XAI, where we examine the relevant literature and address critiques emanating from within the technical sphere. This dual review acts as our foundational anchor, meticulously mapping the historical trajectory, central discourses, and momentous breakthroughs across both arenas. By weaving together insights from AI scholarship with overarching STS narratives, we aspire to identify existing lacunae, intersections, and potential corridors for cross-disciplinary exploration. Following the literature survey, we synthesize the findings, and we elucidate them via pertinent case studies and empirical investigations, where applicable, in order to highlight real-world manifestations and ramifications of explainable AI. These case studies, besides offering tangible illustrations, facilitate the seamless merging of theoretical conclusions with on-ground realities. In this intellectual journey, our navigation will be continually aided by rigorous reflection and immersion within the STS paradigm. With this thesis, we try to furnish a multi-faceted, in-depth viewpoint on explainable AI, harmonizing both its technical nuances and societal implications, thereby charting a roadmap for more ethically sound and informed AI endeavors in the future.

Chapter 2

Methodology

Examining explainable AI through the lens of STS necessitates an interdisciplinary framework. In this endeavor, our research synthesizes insights from secondary literature emphasizing the STS viewpoint, while also drawing from primary sources rooted in the Computer Science and Engineering domains.

For the secondary literature analysis undertaken in this research, we embarked on a systematic approach to gather and analyze scholarly works that intersect the realms of Science, Technology, and Society - Science and Technology Studies (STS) with Artificial Intelligence (AI). Our initial repository of literature was drawn from recommendations made during our MSc program sessions and from invaluable suggestions made by our supervisors. These recommendations provided foundational insights into the domain of STS and AI. Subsequently, to cast a broader net and to ensure comprehensive coverage, we delved into prominent scholarly databases and publishers. These included renowned platforms like JSTOR, ScienceDirect, Elsevier, SAGE, and Springer. Particular attention was directed towards esteemed journals in the field such as “AI & Society”, “Social Studies of Science”, and “Science, Technology, & Society”. To tailor our searches and pinpoint relevant articles, we employed a range of specific keywords. These encompassed terms like “explainable AI”, “AI explainability”, “trustworthy AI”, “AI transparency”, “interpretable AI”, “opaque AI”, and “black-box AI”. These terms were instrumental in ensuring that our research captured the breadth and depth of discussions around AI’s explainability within the STS context. The literature review process was iterative and meticulously structured. In the initial phase, we identified 26 potential articles based on their relevance, as discerned from their titles and abstracts. Upon a cursory review, this collection was refined to include 20 articles deemed particularly pertinent to our research focus. During our meticulous review of these articles, we identified three additional papers from their cited references that appeared germane to our research objectives, bringing our total corpus to 23 articles. Additionally, we also referenced three seminal books in the STS domain. These texts provide not only a comprehensive understanding of the societal dimensions of AI but also delve into the ethical considerations, biases inherent in algorithmic design, and the potential pitfalls of over-relying on technological solutions. They further highlight the intertwined relationship between technology and societal norms, emphasizing the profound influence AI can exert on various societal structures. Through these works, we gain a broader perspective on the multifaceted impacts and challenges of integrating AI into our social fabric.

Regarding the primary literature, the research benefited from the author’s deep knowledge and understanding of the area, accumulated through extensive study as part of the ongoing Ph.D. research Dervakos et al. (2022); Liartis et al. (2021, 2023); Lymperaiou et al. (2022). Throughout the years, the author has extensively engaged with relevant literature related to the broader research topic, which provided him with a strong foundation for understanding the nuances of AI technologies and their sociotechnical implications. Relevant engineering and computer science conferences and journals were also explored to gather insights from the technical aspects of explainable AI and to enrich the existing literature that the author had gathered. Prominent conferences in the field, such as the Association for the Advancement of Artificial Intelligence (AAAI) Conference, the Conference on Neural Information Processing Systems (NeurIPS), the International Conference on Machine Learning (ICML), the Association for Computational Linguistics (ACL) Conference, and the International Joint Conference on Artificial Intelligence (IJCAI) were considered. Furthermore, respected journals including the Journal of Artificial Intelligence Research (JAIR), the Artificial Intelligence Journal, the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), and the ACM Transactions on Intelligent Systems and Technology (TIST) were consulted so as to gain insights into the latest advancements in explainable AI algorithms and techniques. The primary literature incorporated in this thesis draws upon seminal works in the field of AI explainability, AI ethics, social implications of technology, and responsible AI development.

The combined analysis of secondary literature from STS scholarly databases and the incorporation of primary literature accumulated throughout the PhD research, alongside insights from engineering and computer science conferences and journals, allowed for a comprehensive exploration of the topic from both sociotechnical and technical perspectives.

In summary, the methodology for this thesis involved a systematic search and analysis of secondary literature using STS scholarly databases, along with the utilization of primary literature gained through the author's extensive knowledge and understanding of the research area as well as extensive search of prestigious AI-related conferences and journals. This ensured a holistic approach to understanding explainable AI and its aim was to provide a comprehensive and well-rounded exploration of the topic, considering its sociotechnical implications and technical advancements.

Chapter 3

Review and Synthesis of the Secondary Literature

In this chapter, we embark on a comprehensive review and synthesis of the secondary literature in the field of Science, Technology, Society–Science, and Technology Studies (STS) as it pertains to explainable artificial intelligence, and AI in general. The incorporation of an STS perspective in the examination of XAI is of utmost significance, as it enables a nuanced understanding of the socio-technical dimensions, ethical implications, and cultural impacts of AI technologies. While XAI works have shed light on technical aspects and human-centric explanations, an STS approach brings a broader sociological lens to the discourse. It recognizes that the development, deployment, and impact of AI systems are deeply intertwined with social, cultural, and political factors. By engaging with seminal works in STS and papers that explore AI through an STS lens, we aim to shed light on the critical role that this interdisciplinary approach plays in illuminating the complex interplay between AI and society.

It is imperative to acknowledge the extensive body of literature within the STS domain that addresses algorithms, their governance, and consequential societal impacts. While the focus of this section is predominantly on literature specific to artificial intelligence or intelligent agents, it is essential to reference "The Social Power of Algorithms" (Beer, 2019). Authored by David Beer, this foundational work offers invaluable insights into the convergence of STS studies and algorithmic theories. The conceptual frameworks and perspectives elucidated by Beer will significantly inform and shape the subsequent discussions in this study. Beer argues for the underrecognized social power that the algorithms encapsulate as entities/technicalities as does the notion of the algorithm itself. The paper claims that the general conception of algorithms as distant and detached actors prevents us from fully comprehending their social role, leading to what according to Pasquale's terminology is dubbed a "*black box society*", a society "*populated by enigmatic technologies*". The author is concerned about how algorithms relate to power, and he studies different aspects of the matter, including human and machine agency (who and how acts/makes decisions?), the role of algorithms in decision-making ("*how people are treated and judged*"), and the influence of algorithms on the norms and notions of normality and abnormality ("*shaping what they (people) know, who they know, what they discover, and what they experience*"). Regarding agency (a term that the author himself questions whether it is the appropriate one), the article argues that we have the idea that algorithms "*carry some form of agentic power*", thus detaching the algorithm from its human actors, and concludes that it is important to perceive how algorithms make decisions. On the decision-making issue, the author claims that algorithms eventually influence everyday life through their major part in decision-making at many different levels, such as governance. The manuscript also discusses the power of the notion of algorithm and, using Foucault as a springboard, its linkage to truths that it may produce or adhere to. A phrase that sums up this part of the article is "the algorithm exists not just in code but it also exists in the social consciousness as a concept or term that is frequently used to stand for something (something that is not necessarily that code itself)". This is an important distinction between the algorithm as a "physical" entity (if we can call code a physical entity), and as a concept/an idea, that may be a good starting point for studying the algorithm as an abstraction, and not materiality. We will see that Beer's contribution is pertinent throughout the majority of the literature review that follows, setting the foundation for an in-depth study of the field of STS and AI.

To guide our exploration, this chapter is organized into three main sections: (1) STS & AI, (2) STS Approaches to XAI, and (3) Key Themes and Findings from the STS Perspective. Through this structured examination, we aim to provide a comprehensive overview of the rich insights offered by STS scholars and highlight the key themes and findings that emerge when AI is examined through an STS lens.

3.1 STS & AI

AI is a highly complex and multifaceted field that encompasses diverse dimensions. From a technical standpoint, AI involves a range of algorithms, models, and computational techniques that enable machines to perform intelligent tasks. However, AI's significance extends beyond mere technical capabilities. It is intertwined with social, economic, ethical, and cultural aspects, affecting various domains such as healthcare, finance, transportation, and governance. The multifaceted nature of AI lies in its potential to transform decision-making processes, reshape labor markets, raise ethical considerations, challenge notions of accountability and transparency, and influence social dynamics. As such, exploring AI from an STS perspective helps unravel the intricate entanglements of technology and society, providing a comprehensive understanding of its multifaceted nature and its implications in different contexts. Simos et al. (2022) explores this multifaceted nature of AI through a historical study of AI, focused on the electronic era of computing. Its authors argue for "*the existence of two periods; a first period, defined by the discourse of a post-industrial society—and associated notions, like, most notably, "information society"—and a second one, defined by the discourse of another, new, fourth industrial revolution*". Their analysis delves into the intricate political, economic, and social ramifications of AI, providing readers with a comprehensive understanding of its diverse dimensions within a historical framework. By examining the discourses and dynamics between these two distinct periods, their study elucidates the evolving complexities of AI, enabling a deeper comprehension of its varied facets and contextualizing its impact within broader societal contexts.

"Algorithmic Governance" (Katzenbach and Ulbricht, 2019) by Christian Katzenbach and Lena Ulbricht, is another seminal work in the area of algorithms through the STS lens, which explores the emerging field of algorithmic governance, focusing on the increasing influence of algorithms and automated decision-making systems in governance processes, touching upon AI-related matters. It investigates the implications, challenges, and opportunities associated with the adoption of algorithmic systems in diverse policy domains, addressing issues such as accountability, transparency, and the power dynamics between human decision-makers and algorithmic technologies. Their work provides valuable insights into the intersection of governance and algorithms, shedding light on the complexities and implications of algorithmic decision-making in contemporary societies.

In an intricate exploration of the notion of intelligence, the study ”The Smartness Mandate: Notes toward a Critique” (Halpern et al., 2017) offers a pivotal contribution to our research domain by interrogating and re-contextualizing prevailing understandings of smartness and intelligence. It studies the increasing emphasis on incorporating smart technologies and data-driven approaches in various aspects of society, such as cities, governance, and public services. The article questions the tendency that "everything must be smart", it discusses the terminology used when portraying new technologies as "smart" and how this influences modern society. The authors posit that the underlying assumptions and objectives associated with "smart" technologies have gained considerable traction in global policy dialogues. Such acceptance has subsequently catalyzed the development of innovative infrastructures that influence various facets of contemporary life. This smartness mandate draws on multiple and intersecting discourses, with technologies, instruments, and apparatuses binding and bridging these discourses. Throughout their work, Halpern et al. attempt to answer four key questions: "*Where does smartness happen?*", "*What is the agent of smartness?*", "*What is the key operation of smartness?*", and "*What is the purported result of smartness?*". While their analysis of all four questions provides intriguing insights, we will refrain from further exploration in order to maintain relevance to our primary topic. However, we will quote a portion of their discussion pertaining to the agent of smartness:"*Smartness is located neither in the source (producer) nor the destination (consumer) of a good such as a smartphone but is the outcome of the algorithmic manipulation of billions of traces left by thousands, millions, or even billions of individual users.*" Understanding the construction and attribution of "smartness" in technology requires examining the broader social, economic, and political dimensions surrounding the generation and utilization of data. By highlighting that smartness emerges through the algorithmic manipulation of user traces, the authors draw attention to issues of data collection, privacy, and the role of algorithms in shaping the experiences and outcomes of individual users. They prompt critical inquiry into the implications of this process, including questions of control, agency, and the potential for inequalities or biases embedded within algorithmic decision-making.

Meredith Broussard with their thought-provoking works approaches AI, biases, and policy-making from an STS perspective. "Artificial unintelligence: How computers misunderstand the world" (Broussard, 2018) challenges the prevailing narrative surrounding artificial intelligence and highlights the limitations and biases embedded in AI systems. The book adds to the growing literature exploring the limits of AI and techno-solutionism, furthermore showing how its socially-constructed nature replicates existing structural inequalities. Broussard argues that AI is not as intelligent as it is often portrayed and that it can often lead to misguided or incorrect interpretations of the world. This book grounds the sociological analysis in an accessible technical account of the key computational processes involved in machine learning. The computational limitations of machine learning and the absurdity of trying to encode "intelligence" into machines are thereby made more understandable to non-technical readers. Broussard offers a convincing case against what they call ’technochauvinism’, the belief that technology, and in particular AI, can solve everything. We can detect characteristics of the smartness mandate discussed above in Broussard’s analysis, where they claim that for some problems that are being tackled with complex algorithms and AI systems, there may be a better and more obvious low-tech solution. Regarding the autonomy of AI systems, the author emphasizes the need for ‘human-in-the-loop’ systems where computers work in sync with, and augment the capabilities of, humans. Human judgment, they argue, remains essential for dealing with the edge-cases that computers are fundamentally incapable of resolving. An important contribution of this book that is also related to the topic of this thesis, is the discussion about algorithmic accountability. Issues of trust and accountability are strictly linked to the transparency and interpretability of algorithms. The author acknowledges that there have been moves toward a greater understanding of algorithmic inequality and accountability, as well as an understanding of potential violations of privacy in large data sets, and discusses related works, conferences, and labs that have an important contribution.

In their recent work “More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech” (Broussard, 2023), Broussard argues that we are consistently too eager to apply artificial intelligence to social problems in inappropriate and damaging ways. In this book, Broussard again decries technochauvinism and claims that using technical tools to address social problems without considering race, gender, and ability can cause immense harm. The author contends that AI systems often manifest racist tendencies because they are trained on data that mirrors prejudiced actions or policies. Furthermore, these systems are typically developed and tested by individuals belonging to specific demographics, such as able-bodied, white, cis-gender, American men, whose perspectives and experiences may be shaped by their limited exposure to a particular social and cultural context. The book analyzes study cases in various fields of AI such as facial recognition, learning assessment, and medical diagnosis. Racial bias is blatant in some of these cases, for example when facial recognition programs are instituted in policing, leading to harassment and false arrests. Broussard explains how data sets limit the efficacy of AI in predictive policing, as well as in medical diagnostics: “The skin cancer AIs are likely to work only on light skin because that’s what is in the training data.” This work delves deeper into matters of structural and social biases of AI systems through real-life study cases, highlighting the necessity for a change in the way that we design, implement, and train these algorithms toward a more transparent and fair AI.

Cathy O’Neil’s book “Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy” (O’neil, 2017) explores the impact of algorithms and big data on society, particularly in relation to perpetuating inequality and undermining democratic processes. They raise concerns about the unregulated and often opaque use of algorithms in various domains, such as education, employment, criminal justice, and finance. O’Neil argues that there are three elements of a “Weapon of Math Destruction” (WMD): Opacity, Scale, and Damage. They are frequently shielded from public scrutiny, operating as black boxes. Their widespread deployment impacts numerous individuals, thereby amplifying the potential for erroneous outcomes affecting a subset of the population. Furthermore, these systems can engender detrimental effects, such as the perpetuation of biases or enabling predatory practices through selective advertising, and in extreme cases, even contributing to global financial crises. WMD can reinforce biases, discriminate against marginalized groups, and amplify existing inequalities. Hence, the book calls for greater transparency, accountability, and ethical considerations in the development and deployment of algorithms to ensure fairness and mitigate the harmful consequences of unchecked mathematical models.

Jenna Burrell's "How the machine 'thinks': Understanding opacity in machine learning algorithms" Burrell (2016) is another seminal work that is closely related to our topic of research. Burrell investigates how opacity in machine learning can impact decision-making processes, perpetuate biases, and raise ethical concerns. They identify three main forms of opacity, namely "*(1) opacity as intentional corporate or state secrecy, (2) opacity as technical illiteracy, and (3) an opacity that arises from the characteristics of machine learning algorithms and the scale required to apply them usefully*". The second and third forms of opacity are tightly linked to explainable AI since making AI systems understandable to the general public and eliminating the opacity that stems from the complexity of AI models constitute two major objectives of XAI. This taxonomy of the forms of opacity along with Burrell's analysis of each one of them is an important supply for XAI researchers in order to better define the objective and target audience of their methods as well as to improve the design of their systems.

All the works presented above share some common concerns about AI design, creation, and application. Others through the analysis of abstract concepts like smartness and others through real-life examples of biased AI systems, all approaching artificial intelligence from an STS perspective, illuminate the problematic nature of algorithmic black-boxing, particularly in the context of AI. They highlight the necessity for transparency and interpretability as crucial principles to address the inherent challenges and potential biases associated with AI technologies. Through their diverse approaches, these works underscore the importance of critically examining and addressing the complexities of AI within the STS framework and set the foundations for an interdisciplinary approach to explainable AI.

3.2 STS Approaches to XAI and AI Transparency

In our prior discussions, we traversed the broad landscape wherein STS is intertwined with AI. Now, our focal point sharpens as we zero in on Explainable Artificial Intelligence (XAI) and AI transparency, a domain where the confluence of technical design and sociotechnical implications becomes especially pronounced. The pressing need for transparency, accountability, and understanding in AI systems introduces an array of challenges and opportunities that are best examined through an STS lens.

To offer a holistic view of this intricate tapestry, we embarked on a comprehensive literature search. By combing through esteemed STS journals and delving into scholarly databases, as detailed in Chapter 2, we collected a set of articles that shed light on the various dimensions of XAI. What follows is not just a mere presentation of these articles, but a carefully curated synthesis, emphasizing those that hold significant relevance to our central discourse. Through this exploration, we aim to bridge the often-isolated realms of technology development and societal critique. Each article unravels layers of understanding about XAI, ranging from its design ethos to its real-world ramifications, and from the intricacies of its operation to the broader narratives surrounding it.

In "15 challenges for AI: or what AI (currently) can't do" (Hagendorff and Wezel, 2020), Hagendorff and Wezel delineate 15 pivotal challenges poised to shape the future trajectory of AI research. This study categorizes the challenges into three distinct domains: Methodological, Societal, and Technological. Notably, explainability is identified as one of the 15 challenges for AI (namely "Challenge 11: AI applications often lack explainability") emphasizing the intricate complexity associated with deep learning algorithms. The authors critically address the prevalent issue of algorithmic black-boxing and advocate for enhanced transparency within AI systems. Interestingly, several of the challenges identified across the three domains can be adequately addressed by dint of advances in XAI. For instance, XAI methodologies could be instrumental in addressing "Challenge 1: the data AI systems use do not correspond with reality" by elucidating discrepancies between training datasets and real-world applications. Moreover, the pivotal role of XAI is already acknowledged in addressing "Challenge 7: the success of AI applications is tied to their acceptability in society" by fostering trust and enhancing the societal acceptance of AI technologies by both domain experts and the wider public. Such insights underscore XAI's versatile nature. It not only operates as a cross-cutting research domain within AI but also as a foundational pillar with the potential to respond to, at least to some extent, a broad spectrum of challenges that contemporary AI contends with. XAI's inherent attributes extend beyond mere technical clarity; its potential lies in bridging gaps between technology and society. By enhancing the comprehensibility of AI systems, XAI paves the way for more informed decision-making processes, fostering inclusivity by making AI more accessible to various stakeholders. In this context, XAI's role is not merely that of a supplementary tool but also that of an essential component ensuring that AI technologies are developed and deployed responsibly, ethically, and with a clear understanding of their implications. The challenges identified by Hagendorff and Wezel although not directly discussed in their article, serve as a testament to the multifaceted benefits of XAI, emphasizing its potential to contribute holistically to the evolution of AI, addressing concerns from technical hurdles to societal ramifications.

The integral role of XAI in fostering accountability within AI systems is profoundly articulated in Novelli et al. (2023). In this work, the authors posit accountability in AI as a foundational pillar for its governance. The study endeavors to elucidate the concept of accountability, systematically dissecting its architectural components (which include, but are not limited to, the context (purpose of accountability), range (the scope of scrutiny), agent (entity being held accountable), and forum (the audience or entity to which accountability is owed)), and defining accountability goals in terms of the features mentioned above. Within this framework, XAI emerges as a salient mechanism to realize what the authors designate as the "report" goal. This goal emphasizes the imperative to meticulously document the agent's actions and rationales, facilitating a clear exposition and justification to the relevant forum. In essence, this ensures that there exists a comprehensive repository of information that can be utilized to critically assess and potentially challenge the agent's actions, as delineated in the paper. In their concluding remarks, Novelli and colleagues underscore the necessity of addressing the inherent opacity of AI systems, positioning it as a pivotal step in truly realizing accountability in the domain of artificial intelligence.

In Papagni et al. (2023), a comprehensive examination is undertaken to elucidate the intricate dynamics governing the interplay between artificial agents and humans, with a particular emphasis on trust. A central objective of the study is to elucidate the multifaceted relationship between explainability and trust within the context of varying phases throughout an interaction. This paper explores the role of explainability in fostering trust during human-agent interactions, emphasizing the perspectives of non-expert end-users. It delves into the nuanced relationship between trust, reliability, and confidence, accentuating the inherent risks and uncertainties in trust-based dynamics, especially during initial interactions or unexpected events. A comprehensive model is presented, illustrating the interplay between trust, mental model calibration, and explanations across various interaction stages. This work stands as a pivotal reference, possessing significant potential to inform and guide subsequent scholarly endeavors within the realms of Explainable Artificial Intelligence and trust dynamics.

A burgeoning body of literature meticulously examines AI transparency, elucidating its multifarious dimensions and implications. These scholarly contributions approach the topic from a plethora of perspectives, offering rich insights that collectively underscore the complexity and significance of transparency within artificial intelligence paradigms.

In Hollanek (2020) the discourse surrounding AI transparency transcends the conventional bounds of engineering and technicality. Contrary to the prevalent studies that frame transparency as a design-centric issue, the author posits that genuine transparency emerges from incisive critique — an analytical approach aimed at unveiling and challenging entrenched power structures. This perspective argues for a shift from a narrow focus on the operational intricacies of AI systems to a broader contemplation of their foundational premises. Such an interrogation goes beyond the mere mechanics of algorithms, prompting an exploration of the larger socio-economic drivers influencing our engagement with digital entities and the vested interests they cater to. Essentially, Hollanek advocates for an examination "beyond" the technological artifact, striving to discern its broader societal ramifications.

"AI and the expert; a blueprint for the ethical use of opaque AI" (Ross, 2022) by Amber Ross offers a nuanced exploration of AI transparency. While Ross acknowledges the undoubted significance of transparency in AI systems, the work proffers the proposition that alongside championing transparency, the discourse must concurrently address the ethical parameters surrounding the use of non-transparent, or "opaque", AI. The author embarks on an analytical journey, delineating the benchmarks of transparency and reasoning explainability we anticipate from artificial systems, and juxtaposing these expectations with those set for human experts. Central to the discourse is the interrogation of the conditions under which opacity in AI might be ethically tenable: When is such opacity justifiable? Why might it be considered acceptable? And to what extent should this opacity be tolerated? Ross postulates that certain opaque AI paradigms might analogously mirror the dynamics inherent in the relationship between laypersons and specialized experts. Given that experts, in numerous fields, may occasionally proffer judgments rooted in reasoning that might elude the grasp of the general populace, the exploration of such human-centric opacity could illuminate our understanding of, and stance toward, analogous opacities in AI. The author claims that transparency is often used as a proxy for other desired characteristics such as trustworthiness, so we may be ethically permitted to utilize opaque AI technology provided that this trust and trustworthiness can be established through alternate means. The article offers significant insight into the intricate interplay between transparency, trust, and trustworthiness. The analogy drawn between AI and human experts serves as a foundational perspective, inviting a re-examination of AI ethics. However, it is imperative to acknowledge that, based on societal benchmarks established over time, AI may not yet qualify as an "expert" in the traditional sense.

In a vein analogous to the exploration of human-machine transparency standards as delineated by Ross, Zerilli et al. in their work "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" (Zerilli et al., 2019) critically engage with the prevailing notions of transparency and explainability within the realm of algorithmic governance. While they concede the importance of these concepts, they express concern that automated decision-making processes might be subjected to overly stringent transparency standards, potentially stemming from an overestimation of the transparency exhibited by human decision-makers. Drawing upon Daniel Dennet's theory on the "Intentional Stance" (Dennett, 1989), they postulate that just as human action justification aligns with intentional stance explanations, algorithmic decisions should be elucidated in a parallel manner. Operationally, this implies a predilection for explanations of algorithmic decisions that resonate with intentional stance narratives, as opposed to in-depth explorations of the underlying architectural mechanics. In their scrutiny of the potential asymmetry in transparency expectations between human and algorithmic agents, they dissect the primary arguments endorsing such double standards, ultimately concluding that such differential benchmarks lack substantive justification. Günther and Kasirzadeh (2022) present a counterargument to the assertions made by Zerilli et al. (2019) regarding the justification of dual transparency standards. They put forth two cogent arguments, delineating specific scenarios wherein algorithmic decisions necessitate heightened transparency standards in comparison to their human counterparts. Günther and Kasirzadeh present a counterargument to the proposition posited by Zerilli et al. that prioritizes intentional stance explanations over design-centric ones. Drawing upon the tragic case of the Boeing 737 Max 8 aircraft crash during Lion Air Flight 610, attributed to a sensor malfunction, they juxtapose intentional with design stance explanations. Through this analysis, they illustrate instances where the intentional stance is inadequate, advocating instead for design explanations. Additionally, contrary to Zerilli et al.'s assertion that human action justification predominantly aligns with the intentional stance, Günther and Kasirzadeh introduce the symptomatology of Coprolalia—stemming from a specific neurological disorder—as an exemplar. They argue that a purely intentional stance falls short, emphasizing the imperative to incorporate the neurological condition, which they categorize under the design stance, for a comprehensive understanding of the behavior. Drawing from the previously discussed examples, Günther and Kasirzadeh postulate that while design explanations bear significance for both human and algorithmic decision processes, furnishing a comprehensive design explanation for human decisions often proves elusive. Given our innate responsibility in shaping algorithms, compared to our limited influence over human design, they contend that imposing heightened scrutiny on algorithmic decision-making is both warranted and necessary. This heightened scrutiny effectively institutionalizes a default dual standard. Delving deeper, the authors illustrate the intricacies of a Deep Neural Network (DNN) purposed for image recognition. They posit that certain "black box" instances may defy the conventional intentional stance strategy, as our epistemic access to the authentic 'beliefs' and 'desires' governing such algorithms remains restricted. In conclusion, Günther and Kasirzadeh assert that the advocacy for dual standards is underpinned by the predominance of algorithmic design explanations, coupled with the limitations of the intentional stance when confronted with intrinsically opaque algorithms.

In the work of de Fine Licht and de Fine Licht (de Fine Licht and de Fine Licht, 2020), the significance of transparency is analyzed in respect to its influence on the perceptions of the general populace—those directly impacted by AI. Specifically, the study delves into the manner in which transparency, associated with AI decision-making processes, shapes public perceptions regarding the authenticity and acceptability of AI-informed decisions. The researchers postulate that a circumscribed form of transparency, which centers primarily on offering justifications for decisions, might suffice in establishing a foundation for the perceived authenticity of AI decision-making processes, and that full transparency might sometimes even be harmful. In their comprehensive analysis, they critically assess the following major arguments in favor of absolute transparency:

- The assertion that transparency inherently instills a consciousness in decision-makers of their actions being under public scrutiny, consequently emphasizing their obligation to prioritize collective welfare over personal gains.
- The premise suggesting that full transparency augments public comprehension of decisions and the underlying mechanisms, leading to enhanced confidence in the decision-makers.
- The belief that full transparency amplifies the sense of perceived authenticity, as it fosters a sentiment of control within the populace.
- The proposition that transparency generates positive results regarding perceived legitimacy because the public deems the decision-making protocols as fair. This perception subsequently influences their evaluations of both the decisions and of the individuals involved in the decision-making process.

The insights elucidated in this study are both profound and compelling. While the prevailing sentiment in the AI community tends to favor complete transparency as the gold standard, this research offers a paradigm shift, suggesting that a nuanced, selective approach might be more effective in garnering trust and understanding. From our perspective, this paper illuminates potential oversights in current XAI methodologies, where indiscriminate data disclosure is often mistaken for meaningful explanation. This underscores the need for discernment in distinguishing between mere information and genuine elucidative content concerning AI systems.

In a comprehensive examination of transparency in AI systems, Walmsley (2021) postulates that, in scenarios where transparency remains elusive, the principle of contestability may serve as a viable and efficacious substitute. This posits that even in the absence of comprehensive comprehension regarding an AI system's decision-making process, stakeholders retain the capability to challenge and contest its decisions. Initiating the discourse with historical and technological imperatives that catalyzed the contemporary discussions surrounding AI transparency ("*why now?*"), Joel Walmsley delves into the intrinsic significance of the topic ("*why care?*"). Extending this narrative, Walmsley introduces a systematic taxonomy of AI transparency. This categorization divides transparency into two predominant domains: "*outward*" and "*functional*" transparency. The former pertains to the interrelations between the AI system and its extrinsic stakeholders (namely developers, users, and the media), while the latter addresses the intricate internal operations of the AI system. Drawing from this foundational framework, the paper expounds upon its principal assertion: the potential of contestability as an alternative to transparency. Echoing the sentiments presented by de Fine Licht and de Fine Licht (2020), this work censures the dominant perspective that champions absolute transparency as the gold standard. Instead, it underscores the potential of alternative mechanisms, such as contestability, as conduits to achieve the objectives traditionally associated with transparency.

From an angle that diverges yet parallels Walmsley's work in the discussion of contestability, Henin and Le Métayer (2021) contends that while explainability is instrumental, it remains insufficient in ensuring the legitimacy of algorithmic decision systems. The authors advance the hypothesis that justifiability and contestability are the pivotal requirements for high-stakes decision systems. Delineating their conceptual framework, the authors differentiate amongst an explanation, a justification, and a contestation based on their respective objectives. As stipulated in the manuscript, "The goal of an explanation is to make it possible for a human being to *understand* (...), the goal of a justification is to convince that a decision is *good*, while (...) the goal of a contestation is to convince that the decision is *bad*" (emphasis retained from the source). Drawing from this distinction, contestation emerges as the counterpoint to justification, with both deriving from normative frameworks—be they legal, social, or ethical. Conversely, explanations are characterized as inherently descriptive and intrinsically contingent upon the system in isolation. The authors discuss the operationalization of justifiability and contestability while clearly acknowledging that justifiability and contestability cannot be reduced to technical issues. This work exhibits a synergy with preceding studies discussed within this subsection. For instance, while Walmsley (2021) championed contestability as an alternative in scenarios where explainability proves elusive, Henin and Le Métayer propound justifiability and contestability as integral augmentations to explainability to realize overarching objectives, such as accountability and legitimacy. Concepts analogous to justification and contestation have surfaced in the preceding literature, although articulated divergently. For example, Hollanek (2020) alluded to YouTube's video recommender system, utilizing the term "algotransparency"—a construct bearing resemblance to justifications/contestations. According to Hollanek, algotransparency endeavors to decipher the rationale behind YouTube's proprietary recommendation algorithm, not through introspecting the system's concealed logic, but by analyzing its manifest outcomes. This empirical endeavor illuminates the focal areas of scrutiny, gravitating less towards the operational mechanisms of YouTube, and more towards discerning the underlying rationale for its functionality. Such inquiries transcend mere algorithmic intricacies to probe broader forces guiding our digital consumption and the vested interests they cater to. This delineation suggests that algotransparency transcends the fundamental descriptiveness of explanations, contemplating societal and economic norms that either justify or contest the outcomes of such recommendation systems.

Andrada et al. (2023) is another work that stands as a significant contribution to the discourse on transparency in technology, as it discusses varieties of transparency. This study delineates a comprehensive taxonomy, categorizing distinct facets of transparency. This classification serves as a robust framework, facilitating a nuanced understanding of diverse domains within human-technology interactions. Furthermore, the research offers insightful discussions on the nexus between technological transparency and human agency.

While many research endeavors within the realm of Explainable Artificial Intelligence (XAI) often exhibit ambiguous objectives and targets, the inherent value of XAI is predominantly presupposed across the literature. In a nuanced examination, Colaner (2022) delves deeper by probing the intrinsic merit underpinning XAI. Nathan Colaner posits that beyond the instrumental advantages of XAI—such as fostering fairness, trust, accountability, or governance—the provision of explanations holds intrinsic value. This is grounded on the contention that inscrutable systems can precipitate dehumanizing effects. The paper meticulously investigates the potential poseddamages inflicted by these opaque models on individual human dignity, buttressing this analysis with cogent arguments anchored in the principles of participation, knowledge, and self-actualization.

Numerous scholarly articles address the concepts of AI transparency and explainability within the framework of black-boxing, grounding their research on the extant literature that, interestingly, may not primarily focus on AI or intelligent agents. Specifically, Lo (2022) embarks on a nuanced exploration aimed at distinguishing the black-box abstraction enveloping intricate computational systems from the inherent opacity characterizing machine-learning models. Lo postulates that the degree of asymmetry in knowledge is more pronounced in the context of complex software systems than in machine-learning models. To elucidate, while software codes present interpretative challenges that only domain-specialized software experts can competently navigate, the semantics of machine-learning models sometimes remain enigmatic even to the data scientists responsible for their configuration and training. Such intrinsic opacity inadvertently re-balances the asymmetry of knowledge between producers and consumers, given the inherent limitations both groups encounter in comprehending the deep intricacies of ML models. Contrary to conventional perspectives, Lo argues against the arguably quixotic endeavor of seeking human-comprehensible rationales underpinning the decision-making processes of ML models. Instead, he posits that this emergent knowledge equilibrium sets the stage for engendering a more democratized sociotechnical landscape.

Works like those by Carabantes (2020) and Innerarity (2021) delve into the realm of black-box AI, contributing to the multifaceted investigation of the issue. In Carabantes' work, there's an in-depth exploration into the challenges posed by AI's inherent opacity. It examines current methodologies employed to address this opacity and evaluates the rationality of delegating tasks in non-transparent agents. The study not only sheds light on the nuances of machine learning and its associated opacity but also prompts critical questions about the practical implications of using such AI systems. As regards the latter, Innerarity's research offers a detailed analysis of various forms of non-transparency, as well as an exploration of the topic of explainable AI. A key takeaway from Innerarity's paper is the acknowledgment that the complexities inherent in certain realities may surpass individual understanding. Instead, a collective approach might be the pathway to truly grasping these intricate systems. It's worth noting that the forms of opacity discussed in Innerarity's work draw parallels with Burrell's categorization of opacity (Burrell, 2016) previously discussed in this section.

A plethora of research projects that focus on transparency and explainability exist within the STS domain. While many of these studies approach the topic from angles distinct from our research focus, their insights remain invaluable. These works facilitate a broader understanding of the societal, cultural, and ethical dimensions underpinning transparency and explainability, offering a comprehensive framework to contextualize our specific observations and conclusions. While a detailed exploration of all such works is beyond the purview of our study, we highlight a few interesting articles that resonate with our themes. In "Towards Safe AI" (Morales-Forero et al., 2023), the authors embark on a journey to enhance AI safety, striving to bridge existing methodological chasms in system engineering. Utilizing the Box-Jenkins method for statistical modeling, the study discerns potential pitfalls in calibrating and validating AI models, suggesting the best strategic practices to mitigate these challenges. The pivotal role of transparency in achieving AI safety is acknowledged, as well as the instrumental value of explainable AI in fostering such transparency. The paper delves into various extant explainability methodologies and reflects upon their relevance in shaping AI governance and regulatory frameworks. In Ball and Koliousis (2023), Ball and Koliousis elucidate the crucial role philosopher engineers play across three foundational design stages of AI systems: deployment management, objective setting, and training. Through illustrative examples, they exemplify the potential of philosopher engineers in fostering equitable decision-making in AI, even when contending with intrinsically biased training data. The authors further sketch their vision of an interdisciplinary educational paradigm to cultivate this breed of AI professionals. Their perspectives underscore the profound contributions philosopher engineers can make in the realm of explainable AI, stressing the indispensable nature of a multidisciplinary approach to XAI. Finally, Shin et al. (2022), while not directly aligned with our research focus, offers pivotal insights. Exploring user sensemaking in fairness and transparency within over-the-top platforms, Shin et al. postulate pertinent questions about the intended audience and objectives of XAI systems' explanations. Their advocacy underscores the paramount need for AI systems to be explainable to end-users, raising questions regarding the relevance of information provided in explanations.

3.3 Key Themes and Findings from the STS Perspective

Building on the foundational exploration of STS in relation to AI and XAI as established in preceding subsections, we have undertaken a rigorous analysis of the literature at hand. Through this meticulous review, we have distilled a series of key themes and findings that rise to prominence within the STS framework. This subsection delves into these central insights, offering a synthesized perspective on the nuanced interplay between STS, AI, and XAI, and further elucidating their intertwined roles and ramifications in the broader socio-technical context. The multifaceted political, economic, and social implications of AI emerge as a recurring theme across the vast majority of scholarly works discussed above. Such consistent attention not only accentuates the complex nature of AI's role in society but also drives home the importance of a holistic understanding. It becomes evident that an interdisciplinary approach to AI research is indispensable. This approach ensures that we not only identify the potential and pitfalls of the technology but also appreciate the myriad ways it intersects with, and impacts, societal structures. In understanding these intersections, it becomes clear that AI is not merely a tool, but also an actor influencing and being influenced by societal norms and expectations. Against this backdrop, a prevailing sentiment within the technical community posits AI as a panacea for a host of societal challenges. This notion, termed *technochauvinism* by Broussard, faces rigorous scrutiny in interdisciplinary circles. Such studies shed light on the importance of tempering our expectations, recognizing that while AI offers transformative potential, an over-reliance on it, without considering the intricate socio-cultural dynamics, can be counterproductive. Hence, it is vital to approach AI not just as a solution provider but also as an entity that interacts dynamically with its environment, necessitating a thoughtful and informed approach to its deployment and governance.

Venturing further into the realms of XAI and AI transparency, the significance of XAI becomes evident. It emerges as an instrumental facet in advancing the AI field. XAI is not only a technical enhancement but it also addresses pressing challenges that currently confront AI, particularly in its relationship with society. Key among these challenges are issues of accountability, societal acceptance, and the dual concerns of fostering both trust in and trustworthiness of AI systems. A salient insight from the STS literature emphasizes that XAI extends beyond being a mere technical challenge or a singular design consideration. Instead, AI ought to be understood within the broader power dynamics in which it operates, and XAI should be constructed with this contextual understanding in mind. While traditional narratives often portray explanations as intrinsically descriptive, there is a growing scholarly discourse that underscores the need to move beyond basic descriptiveness. It advocates for the integration of societal, ethical, and legal frameworks in explanations, positioning them as tools to validate or critique the behaviors and outcomes of AI systems.

A significant revelation within the STS literature pertains to the perception of full transparency as an unequivocal benchmark for AI. This universally accepted standard is now subject to critical scrutiny, with numerous studies challenging its supremacy. Interestingly, some studies even allude to scenarios where, beyond its unfeasibility, full transparency might inadvertently cause harm. Although the overarching consensus acknowledges the importance of transparency, a burgeoning sentiment within the academic community posits the need to explore alternatives, especially when achieving transparency proves elusive. This exploration also encompasses potential contexts where the deployment of inherently opaque systems could be deemed acceptable. Certain scholars advocate for a nuanced approach, championing selective transparency as a means to address the multifaceted challenges associated with XAI. Concurrently, there's a discourse on the feasibility of realizing the aspirations of XAI via avenues other than sheer transparency or explainability. Central to this debate is the task of precisely defining and understanding the terms "transparency" and "opacity". To this end, numerous STS-oriented investigations have undertaken the responsibility of dissecting these terms, subsequently presenting taxonomies that delineate their varying dimensions. Such categorization facilitates a granular study of transparency's different facets, encouraging tailored interventions for each specific form of opacity. This intricate exploration stands in stark contrast to the predominant approach within technical literature, whereby transparency often gets painted with a broad brush, either being oversimplified or depicted as a singular, monolithic entity. By thoroughly analyzing the multifarious shades of opacity and transparency, we are better positioned to pinpoint the primary goals of XAI, which in turn can inform and refine the design strategies for explainability mechanisms.

Building upon the discussions of transparency standards in AI, the STS scholarship offers an intriguing comparative analysis between the standards set for machines and those for humans. A segment of this literature posits that the benchmarks established for algorithms might occasionally be more stringent than warranted. Conversely, other scholars propose that machines ought to be held to more rigorous standards than their human counterparts. While there might not be a unanimous consensus on the idea of a double standard between humans and machines, a salient point emerges: the benchmarks for machine transparency can be effectively framed through an examination of human transparency and its legal, ethical, and social acceptability. The endeavor to juxtapose standards for machines and humans, and the subsequent societal reception of each, stands as a pivotal contribution from the STS domain. This perspective elevates the discourse beyond a purely technical realm; it enriches our understanding by drawing upon a vast reservoir of knowledge and research on human behavior and social interactions accumulated over decades across diverse disciplines.

Chapter 4

Presentation of the Primary Literature

In this chapter, we delve into the analysis of primary literature that forms the foundation of this thesis. The primary literature includes seminal works, scholarly articles, and research papers related to the broader research topic of explainable AI, accumulated through extensive study as part of the author's ongoing Ph.D. research. The main focus of this primary literature review is to approach explainable AI from the engineering and computer science perspective, to gain more technical insights on the matter and to facilitate a comprehensive understanding of the topic for a more in-depth study of XAI as a multidisciplinary field of research.

4.1 Overview of the Research Area

The research area under investigation revolves around the concept of explainable AI (XAI) and its sociotechnical implications. Explainable AI has gained significant importance in contemporary society as AI systems become increasingly integrated into various domains. The lack of transparency in AI decision-making has raised concerns about accountability and trust, necessitating the development of explainability mechanisms.

Within the field of explainable AI, several key themes, challenges, and debates have emerged. Ethical implications associated with black box algorithms, including bias, discrimination, and societal inequalities, have drawn considerable attention. Striking a balance between transparency and performance trade-offs remains a challenge. The importance of user trust, regulatory frameworks, and ethical guidelines in AI development and deployment cannot be understated.

Advancements in explainable AI research have led to various approaches and methodologies. Rule-based explanations, model-agnostic methods, and interpretability through visualizations are some of the techniques employed to achieve explainability. Understanding the technical advancements, limitations, trade-offs, and applicability of these approaches is crucial.

However, explainable AI is not solely a technical issue—it is deeply intertwined with sociotechnical dynamics. Exploring the interaction between technology and society is essential. Explainability in AI has implications for different stakeholders, including end-users, developers, policymakers, and the broader public. Considering how AI systems are embedded within sociotechnical systems and influenced by social and cultural factors is crucial. Power dynamics, accountability, and responsibility play a significant role in shaping discussions surrounding explainable AI.

Despite the existence of notable pertinent research contributions, gaps, controversies, and unanswered questions persist within the field. The identification of these research gaps highlights the need for a nuanced understanding of the sociotechnical implications of AI explainability. The primary literature analysis aims to address these gaps and to contribute to the ongoing discourse.

This overview provides a comprehensive understanding of the research area, setting the context for the primary literature analysis. By exploring the sociotechnical dimensions of explainable AI and the challenges associated with achieving transparency and accountability, the primary literature analysis gains a solid foundation. This understanding paves the way for a critical examination of selected works, contributing to the broader understanding of explainable AI from an STS perspective.

4.2 Explainable AI from a technical perspective

Since the early days of Artificial Intelligence, the interpretability of algorithms has been a topic of ongoing discussion and debate. As AI systems became more sophisticated and capable of complex decision-making, questions arose regarding the transparency of their inner workings and the ability to understand and explain the reasoning behind their outputs. Researchers and experts have grappled with the challenge of striking a balance between achieving high-performance AI models and ensuring interpretability, which led to a rich body of research and exploration in the field of explainable AI. In this section, we delve into the technical aspects of explainable AI, exploring the algorithms, methodologies, and techniques used to achieve transparency and interpretability in AI systems. We will commence by providing an introduction to the primary types of AI explanation methods that have been developed to address the need for interpretability in artificial intelligence systems. Subsequently, we will engage in a comprehensive analysis of specific key works in the field, examining their contributions and advancements in the realm of explainable AI. Over the last few years, there have been numerous studies on explainable AI. Here we give a brief overview, and we discuss seminal works and key concepts introduced in recent literature; however, for a comprehensive analysis of the explainability methods that have been developed, we refer readers to Guidotti et al. (2018) and Bodria et al. (2023).

4.2.1 Taxonomy and Methodological Approaches of Explainable AI

While a consensus exists among the majority of XAI studies regarding the fundamental attributes that differentiate between various explanation methods and the principles guiding the categorization of XAI systems, subtle variances can be observed across different scholarly works. In this thesis, we adopt the primary taxonomy and terminology proposed in Guidotti et al. (2018) as the foundational framework, while concurrently reviewing, extending, or adapting it as necessary to ensure a comprehensive analysis of explainability methods in accordance with the research objectives. Below we outline some essential characteristics employed to distinguish XAI methods, along with a discussion of prominent approaches within this domain. By examining these key characteristics and distinct methodological approaches, we aim to provide a comprehensive overview of the landscape of XAI research.

Global and Local Intepretability One crucial differentiating characteristic of explanation methods lies in the scope of the explanations they provide. On the one hand, explanations that offer comprehensive insights into the entirety of an AI system's workings are usually referred to as "*global explanations*". Such explanations enable an understanding of the holistic logic of a model and illuminate the reasoning process leading to all the possible outcomes. On the other hand, "*local explanations*" refer to interpretations that specifically address individual predictions or decisions made by the system under investigation. These explanations focus on providing insights into the rationale behind specific instances or cases, offering a more localized perspective within the broader AI system.

Transparency and Post-hoc Intepretability Transparency and post-hoc interpretability represent two distinct approaches in the quest for understanding and explaining the decisions made by AI systems. On the one hand, transparency refers to the use of inherently interpretable models that provide clear and understandable explanations for their predictions or decisions. Transparent models, such as decision trees, rules, and linear models (Freitas (2014); Huysmans et al. (2011); Ribeiro et al. (2016a)), reveal the factors influencing their outcomes usually through their simple and non-complex structure. On the other hand, post-hoc interpretability techniques focus on generating explanations for black-box models, which lack inherent interpretability. These methods aim to extract insights from the model's behavior by analyzing input-output relationships or perturbing the inputs to understand the model's response. While transparency offers direct interpretability, post-hoc interpretability provides a means to gain insights into complex models that usually exhibit high predictive performance but lack inherent explainability. The choice between transparency and post-hoc interpretability depends on the specific requirements, trade-offs, and constraints of the AI system under investigation.

Black-Box and White-Box Approaches Post-hoc explainability encompasses two primary approaches: black box and white box. Black box approaches refer to methods that focus on interpreting the decisions and predictions of AI models without direct access to their internal mechanisms. These methods treat the AI model as an opaque entity, examining its input-output behavior to generate explanations. Techniques such as feature importance analysis, sensitivity analysis, or model-agnostic methods are employed to gain insights into the model's decision-making process. Black box approaches are particularly useful when dealing with complex models, such as deep neural networks, where the internal workings may be intricate and challenging to interpret, as well as cases where the system under investigation is not accessible (due to copyrights, intellectual property protection, corporate secrecy, etc.). In contrast, white-box approaches encompass direct access to the inner mechanisms of the AI system, allowing for a comprehensive examination of parameters, intermediate representations, and other relevant factors. Through this access, white-box approaches facilitate a deeper understanding and provide additional insights into the underlying behavior of the system. White-box approaches are occasionally referred to as transparent-box methods.

Numerous additional characteristics are used to categorize and classify explanation methods, encompassing factors such as the domain of application (e.g., tabular data, images, text) and the type of explanation produced (e.g., rules, saliency maps). However, here we have outlined key concepts in the domain of explainable artificial intelligence (XAI) that are relevant to this thesis, and provide a foundational understanding for subsequent discussions. For a more comprehensive exploration of the topic, we recommend referring to the following extensive XAI surveys, which provide in-depth insights into the landscape of XAI research and methodologies: Guidotti et al. (2018), Bodria et al. (2023), and Das and Rad (2020).

In addressing the challenge of explaining AI systems, researchers have explored diverse approaches that leverage a range of methods and techniques to generate explanations in different forms and formats. Below we provide a concise overview of several types of approaches employed in the field of AI explainability. This non-exhaustive list aims to offer the reader a glimpse into the diverse range of methodologies used to tackle the challenge of explaining AI systems.

- *Decision Tree Approximation* involves constructing a simplified decision tree that approximates the behavior of the underlying model. By mapping the input features to a sequence of hierarchical decision rules, decision tree approximation offers a transparent and intuitive representation of how the model arrives at its predictions. This approach exemplifies a prevalent strategy employed in the field of explainable artificial intelligence (XAI) known as mime explainers. Mime explainers involve approximating the behavior of an AI system with an interpretable mime, serving as a surrogate model that provides transparent insights into the system's decision-making process.

- *Rule Based Methods* generate explanations in the form of explicit rules, which capture the decision logic and reasoning of the underlying model. Some works argue that rules are among the more human-understandable techniques and might be the desirable form of explanations Guidotti et al. (2018); Pedreschi et al. (2019).

- *Feature importance* is used to identify and quantify the contribution of input features in the decision-making process of a model, providing insights into which features have the most significant impact on the model's predictions or outcomes.

- *Prototypes* involve the identification and characterization of representative instances or examples that encapsulate the decision boundaries and behavior of an AI model.

- *Counterfactual Explanations* present alternative scenarios that could have led to different outcomes. These explanations offer hypothetical changes to input features or conditions, allowing users to understand the causal relationships between inputs and outputs. They basically try to answer the question *"What changes to the input features or conditions would have resulted in a different outcome or prediction?"*.

4.2.2 Seminal XAI Works

This subsection delves into a compilation of significant contributions within the domain of explainable artificial intelligence (XAI). These works recognized as influential and groundbreaking in the field, have played a pivotal role in advancing our understanding of AI interpretability and have laid the foundation for subsequent research and developments. By exploring these seminal works, we aim to highlight the key insights, methodologies, and contributions that have shaped the landscape of XAI, providing valuable perspectives for researchers, practitioners, and stakeholders seeking to delve deeper into the field. Ribeiro et al. (2016a) is considered to be a seminal work in the area of explainable AI. In this paper, the authors introduce LIME (Local Interpretable Model-agnostic Explanations), which is independent of the type of data, and the kind of black box to be opened. The

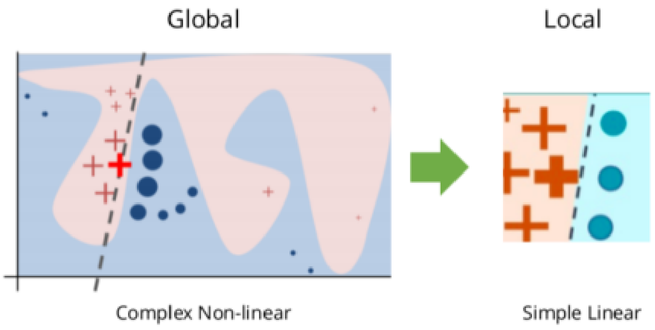

Figure 4.1: LIME's basic principle approximating complex models locally with simpler linear models.

primary idea behind LIME is that the explanation can be produced locally from the records generated randomly in the neighborhood of the record to be explained and weighted based on their proximity. This basic principle of local approximation of a complex model with a simpler linear model is shown in Figure 4.1. The authors use linear models in their research as understandable local predictors that yield the significance of the attributes as an explanation. The necessity of transforming any kind of data into a binary format that is purported to be understandable by humans is a weakness of this strategy. In addition, linear models and the significance of their features are the only ones used in practice to provide the explanation. LIME has been the foundation and inspiration for numerous works that followed the same rationale, approaching the problem by sampling that data space locally around a specific data point and approximating the behavior of the AI system in that neighborhood. Ribeiro et al. (2018) introduces an extension of LIME that uses decision rules as local interpretable classifiers. A bandit algorithm is used by the *Anchor* to generate the anchors with the highest coverage while adhering to a user-specified precision level. An anchor explanation is a decision rule that sufficiently ties a prediction locally so that changes to the other feature values are irrelevant; in other words, similar instances covered by the same anchor have the same prediction outcome.

Another set of works produces saliency masks incorporating network activations into their visualizations. Selvaraju et al. (2017) did some major contributions to that area, introducing Grad-CAM which visualizes the areas of an input image that strongly influence the predictions of a convolutional neural network (CNN). By leveraging the gradients flowing into the final convolutional layer of the CNN, Grad-CAM generates a heat map highlighting the regions that contribute most to the predicted class. This localization technique provides intuitive visual explanations by overlaying the heat map onto the input image, allowing users to understand which parts of the image are critical in influencing the network's decision. Grad-CAM has gained popularity due to its simplicity, effectiveness, and applicability to a wide range of CNN architectures, making it a valuable tool for interpreting and explaining the predictions of deep learning models in image classification tasks. Figure 4.2 shows an example of saliency maps produced with Grad-CAM that indicate the regions that contribute most to the predicted class ("Cat" and"Dog" respectively).

Lundberg and Lee (2017) is another influential work in the area of XAI, where the authors introduce the SHAP (SHapley Additive exPlanations) framework, which provides a unified approach to interpreting predictions from a wide range of machine learning models. The paper presents an

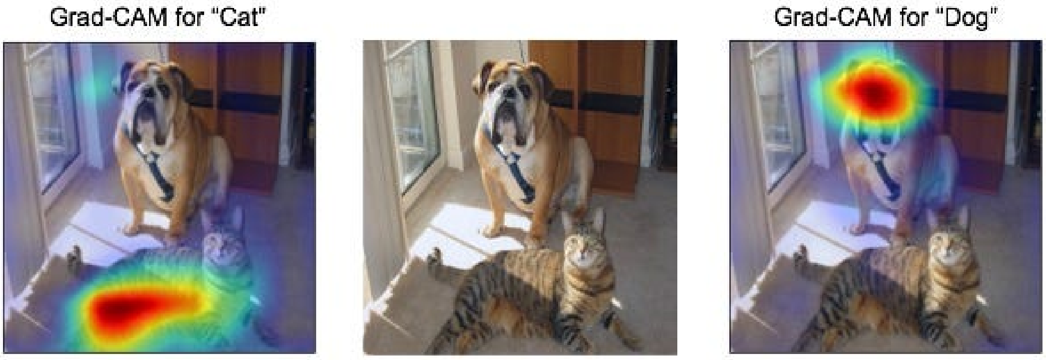

Figure 4.2: An example of saliency maps produced with Grad-CAM from Selvaraju et al. (2017).

explanation technique rooted in cooperative game theory, specifically using Shapley values, to assign feature importance scores to each input feature. By considering all possible feature coalitions and their contributions, SHAP captures the interaction effects and quantifies the impact of each feature on the prediction. This unified approach offers a comprehensive and mathematically grounded methodology for understanding and interpreting model predictions, empowering users to gain insights into the relative importance of input features in the decision-making process.

Counterfactual explanations play a major role in the XAI community and have been established as one of the most popular and intuitive forms of explanations. A typical example often used to showcase a real-life scenario of a counterfactual explanation is the loan application, where an applicant applies for a loan but receives a rejection from the bank. A counterfactual explanation can be used to explore what changes in the application could have led to approval, and which of these changes may be "minimal", and thus easier for the applicant to change. For example, the counterfactual explanation might reveal that their loan application would have been approved if the applicant had a higher credit score, a lower debt-to-income ratio, and a longer employment history. By highlighting these specific factors, the counterfactual explanation provides actionable insights to the applicant, allowing them to understand the areas of improvement required for a successful loan application. It helps applicants make informed decisions and take steps to address the identified deficiencies, increasing their chances of securing a loan in the future. The usability of counterfactual explanations in this context lies in their ability to guide applicants by illustrating the specific changes required to achieve a desired outcome. Contrastive explanations are also closely related to counterfactual explanations (many works also refer to counterfactual explanations as counterfactual contrastive explanations). Although minor differences may exist according to some definitions between contrastive and counterfactual explanations, in most works they are studied as a unified approach, and therefore we will follow the same rationale. Numerous works have been published in this area, providing different technical solutions and areas of application. Some key works include (but of course are not limited to) Wachter et al. (2017) that explores the use of counterfactuals for explainability in the context of automated decision-making under the General Data Protection Regulation (GDPR), and Dhurandhar et al. (2018) that introduced the "pertinent positives" as a factor whose presence is minimally sufficient in justifying the final classification and "pertinent negatives" as a factor whose absence is necessary in asserting the final classification. Many researchers have also focused on the close connection of this type of explanation to human reasoning and how humans produce such explanations in their everyday lives Byrne (2019). Some works have also addressed the feasibility of the changes suggested by counterfactual explanations, since for instance in the loan application example mentioned above, if the applicant was 7 years younger the loan could have been approved, but it is not feasible for the applicant to become younger. Poyiadzi et al. (2020) considers the constraints and feasibility of counterfactual changes, ensuring that the proposed changes are realistic and achievable in real-world scenarios. By providing actionable counterfactual explanations, the approach aims to guide users toward making meaningful changes that can lead to desired outcomes. Recent works have also studied the evaluation of counterfactual explanations, like Filandrianos et al. (2023) that I have co-authored, where an iterative feedback approach is employed to evaluate the minimality of suggested counterfactual edits for language models.

4.3 Critical Perspectives from within the technical community

Although existing XAI methods have made notable strides in enhancing transparency and interpretability, they exhibit certain limitations that call for a critical examination. These limitations extend beyond the technical aspects and encompass their broader societal implications. Existing methods often prioritize technical efficacy without fully considering the needs of end users and the societal impact of AI decision-making. Moreover, XAI has predominantly been approached from a computer science perspective, overlooking the interdisciplinary nature of the field that necessitates insights from social sciences, ethics, and policy as well. Over the last year scholarly works in reputable computer science conferences and journals that offer critical scrutiny of the prevailing approaches in Explainable Artificial Intelligence (XAI) have emerged. These are works from computer scientists, social scientists, and policymakers, who not only challenge the existing methods but also advocate for a change in the way we view and address Explainable AI.

"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead" Rudin (2019) is a seminal work by Cynthia Rudin that (as its title implies) presents a compelling argument against relying on black box models for high-stakes decision-making and advocates for the use of interpretable models instead. Rudin emphasizes the importance of using models that offer clear explanations and understandable decision rules, particularly in domains where human lives, rights, or fairness are at stake. She highlights significant limitations inherent in current approaches within the field of explainable artificial intelligence (XAI). By adopting a broader perspective, the paper takes a critical stance and questions the fundamental principles underlying certain methodologies that purport to offer transparency and explainability in AI systems. This critical examination encourages a reevaluation of prevailing notions and fosters a more holistic understanding of the challenges and requirements for achieving true transparency and explainability in AI systems. This comprehensive analysis holds significant value within the XAI community, serving as a valuable source of motivation and guidance for future research. It fills a critical gap in the field, as evidenced by various studies (see for example the discussion about Lipton's work below), where concerns have been raised about the conflicting and ambiguous motivations and assertions of various XAI methods. One important part of Rudin's work refers to the mime approach, where an interpretable mime approximates the behavior of the complex model. The main claim is that if the mime (explainer) is actually good, we should get rid of the opaque model and use the mime instead, so there is no point in creating mime explainers that utilize the same raw data as the model under investigation does, because they are either not good enough, or the model under investigation is redundant. As discussed in the paper, even if both models are correct (the original black box is correct in its prediction and the explanation model is correct in its approximation of the black box's prediction), it is possible that the explanation leaves out so much information that it may be misleading. Rudin also touches on the matter of terminology used within the academic community in the area of XAI where many methods claim to provide explanations while in fact, they provide a statistical summary of the model's behavior without any guarantees about what the model actually takes into consideration, hence alternative terms like "summary statistics" or "summary of predictions" are proposed to be used instead of the term explanations. Another important aspect of this paper is the question of the often-used argument that there is a trade-off between model performance and interpretability. Rudin claims that this is not necessarily true, especially in cases where structured data are available. Despite being published nearly four years ago, the criticisms raised in this paper remain relevant and applicable to a majority of existing XAI methods. This observation highlights the ongoing need for progress and advancement in the field, indicating that there is still a substantial distance to cover before achieving comprehensive and robust solutions for transparency and explainability in AI systems. The longevity of these criticisms underscores the continued importance of addressing the identified shortcomings to work towards more effective and trustworthy XAI methodologies and showcases the importance of Rudin's work.

Zachary Lipton's "The Mythos of Model Interpretability" Lipton (2018) is another major work in this field. Lipton examines the diverse and sometimes conflicting motivations for seeking interpretability, and he identifies transparency to humans and post-hoc explanations as competing notions. The paper questions commonly made assertions about the interpretability of linear models and deep neural networks, challenging the notion that linear models are inherently interpretable while deep neural networks are not. In this work, Lipton highlights the absence of formal technical meaning of key concepts like interpretability and argues that "*interpretations serve objectives that we deem important but struggle to model formally*". He identifies five key objectives: Trust, Causality, Transferability, Informativeness, and Fair and Ethical Decision-Making. Lipton shows how ambiguous these objectives might sometimes be, and how they are not strictly linked to transparency as it is often portrayed. An important claim of this paper is that interpretability does not constitute a monolithic concept. In order for assertions about interpretability to be meaningful, it is essential to establish a precise definition (which might not be the same in all cases). Presently, a considerable number of works within the field of explainable artificial intelligence (XAI) make claims about interpretability without offering a clear definition of the concept itself, lacking explicit objectives, and failing to consider the perspectives and requirements of end-users.

Tim Miller's "Explanation in artificial intelligence: Insights from the social sciences" Miller (2019) delves deeper into the notion of explanation and argues that leveraging research from philosophy, psychology, and cognitive science can provide a valuable foundation for developing effective explanations in AI. Miller emphasizes the need to move beyond intuitive notions of what constitutes a "good" explanation, and he explores cognitive biases and social expectations that influence the explanation process. The paper, in its four main chapters, discusses four fundamental questions: "*What Is Explanation?*", "*How Do People Explain Behaviour?*", "*How Do People Select and Evaluate Explanations?*", and "*How Do People Communicate Explanations?*". Even by simply glimpsing at these questions, we can see the main role that human perception plays in Miller's analysis. This paper stands out from the existing literature because it tries to define the very problem of explainability by way of a theoretical study of the foundations of explainability, starting from the explanation itself, and this is what makes it a seminal work in the area of XAI. The paper contributes to the creation of a taxonomy for explanations, defining types and levels of explanations based on Aristotle's Four Causes model as well as findings from cognitive science about how humans perceive explanations. It lays a robust groundwork for future research in the area by establishing a common understanding of the fundamental concepts of explainability and advocates for a multidisciplinary perspective for XAI that incorporates insights from the social sciences and emphasizes the communication of explanations. As depicted in Figure 4.3, Miller places XAI within the domain of human-agent interaction, highlighting the crucial role of explanation communication. In this context, an explanation is not only expected to encompass informative details about the behavior of AI systems but also needs to be comprehensible and meaningful to the intended audience. This emphasis on effective communication underscores the significance of tailoring explanations to align with the understanding and relevance of the target audience, ultimately enhancing the interpretability and usability of AI systems in real-world scenarios. Through the use of very simple and intuitive examples, Miller shows how explanations are relevant to the end-user, something that is often ignored in XAI literature where one explanation is usually provided for everyone, without considering the objectives, familiarity and above all relevance to the end-user. For example, in a fatal car accident "*consider how the cause of death might have been set out by the physician as 'multiple haemorrhage', by the barrister as 'negligence on the part of the driver', by the carriage-builder as 'a defect in the brake lock construction', by a civic planner as 'the presence of tall shrubbery at that turning'. None is more true than any of the others, but the particular context of the question makes some explanations more relevant than others.*" Choosing the relevant explanation for the end user can be vital for understandability and trustworthiness. XAI methods need to be able to differentiate according to the end-user in order to be able to provide relevant explanations. For example, in the case of a medical assisting AI system, the doctor that uses the system would expect completely different information regarding the operation of the model, compared to the AI engineer that designed and implemented it. These two end-user groups require different levels of explanations and mixing information from different levels may prove to be misleading and non-understandable. The above also highlights the absence of analysis of the objectives in XAI works, since the end-user and the objectives are tightly connected.

Other works like Adrian Weller's "Challenges for Transparency" Weller (2017) had already underlined the problem of ambiguity of fundamental terms in the area of XAI like transparency. They identify different types and goals of transparency for various audiences and use cases. Interestingly, they include the broader society as a stakeholder that "*needs to understand and become comfortable with the strengths and limitations of the AI system, overcoming a reasonable fear of the unknown*". They also highlight the necessity for a proper evaluation of transparency and ways to measure the interpretability of AI systems. Weller's work also challenges the perception that transparency is de facto good and discusses settings where transparency might actually lead to a worse outcome, raising concerns about privacy, robustness, and deliberate misleading use. Based on earlier works (Langer et al. (1978), Wiley (1983)) they revisit the issue of communication vs. manipulation and discuss how meaningless explanations might achieve the goal of the explanation, without actually

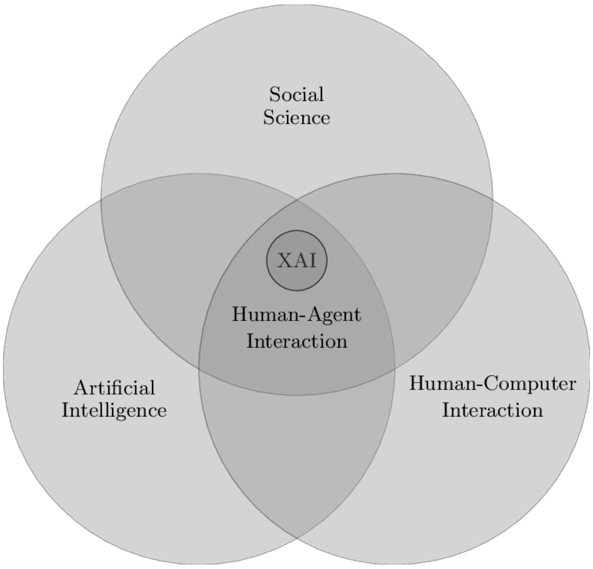

Figure 4.3: XAI as an interdisciplinary field from Miller (2019).

providing useful insights (or sometimes any information at all), supporting their claims for the need of evaluation of explanations. They also reference recent works (Levine and Schweitzer (2015)) that, by employing a similar rationale, have demonstrated scenarios in which prosocial lies can enhance trust, thus illustrating that in certain cases, the mere presence of an explanation holds greater significance than the specific information it conveys.

This subsection has provided a glimpse into a selection of significant works that critique the field of explainable artificial intelligence from within the technical community. These works have raised thought-provoking questions, identified limitations, and stimulated discourse regarding the methodologies, definitions, and objectives of XAI. However, it is important to acknowledge that the scope of this thesis did not encompass an exhaustive review of all relevant works in the XAI domain. Numerous other works exist, each offering unique perspectives and contributing to the ongoing evolution of XAI. Therefore, we encourage readers to explore the extensive body of related literature to gain a comprehensive understanding of the diverse viewpoints and ongoing advancements in the field of XAI.

Chapter 5

Interpretation of Findings: Bridging STS and Technical Insights in XAI

Exploring eXplainable Artificial Intelligence (XAI) through both technical and STS lenses provides a comprehensive and nuanced understanding of the field, its prevailing challenges, and prospective advancements. These dual perspectives, while distinct in their approaches, synergistically enhance our grasp of the subject. The STS perspective delves into the intricacies of AI transparency and XAI, offering a foundational socio-cultural context that can inform and guide technical endeavors. Conversely, technical research in XAI surfaces specific challenges and considerations that may not always be evident from a purely STS standpoint. Recognizing the insights offered by both realms is pivotal. Their convergence not only fosters a holistic view of AI transparency and XAI but also paves the way for more integrative, robust, and culturally attuned technical solutions in the future.

In this chapter, we aim to synthesize and interpret the insights garnered from our comprehensive literature review, placing emphasis on salient findings. To enhance clarity and offer intuitive insights, we'll delve into illustrative case studies. Throughout this chapter, we will anchor our discussions around two primary case studies that serve as foundational examples upon which we can elucidate our findings and interpretations. As we delve deeper into specific topics or nuances, we may introduce supplementary examples or additional case studies to further illustrate and substantiate our points. The first case study explores the workings of a music recommender system, akin to platforms such as Spotify. This represents a proprietary, black-box AI system operating in a non-critical domain. Conversely, the second case study examines an open-source medical assistant designed to aid oncologists. This tool, developed by a public institution like a university, sifts through intricate patient data to propose potential diagnoses and recommend personalized treatment plans for various cancers. Here, we're dealing with a more transparent (according to some definitions and forms of transparency), open-source AI system in a high-stakes domain. By dint of these contrasting examples, we aim to elucidate various facets of transparency and opacity in AI systems, demonstrating how our combined insights from both our primary and secondary literature reviews manifest in real-world contexts.

5.1 Navigating the Lexicon of Explainable AI: A Tangle of Terminology

As the domain of Explainable AI grows in prominence, the clarity and precision of the terminology employed become increasingly vital. A shared and clear lexicon is foundational for facilitating interdisciplinary communication, ensuring robust research methodologies, and for the practical implementation of systems that adhere to XAI principles. Unfortunately, the rapidly expanding research landscape in XAI has witnessed a proliferation of terms, sometimes used interchangeably and often without clear definition, leading to potential confusion and misalignment in objectives. The nuanced distinctions between terms like 'transparency', 'interpretability', and 'explainability', which might seem academic at first, can have profound implications for the design, evaluation, and deployment of AI systems. The inconsistent use of key terms across different studies can muddy the waters, making it challenging to synthesize findings, share insights, and establish common ground. Furthermore, the lack of clear definitions can inadvertently lead to nebulous objectives and an unclear understanding of the intended end-users for XAI systems. This section aims to dissect and clarify the intricate web of terminology within the realm of XAI, highlighting the complexities, inconsistencies, and hidden consequences.

5.1.1 The Elusive Clarity of AI Transparency

Transparency, in the context of Artificial Intelligence (AI), has become a pivotal term. But what does it truly mean to have a transparent AI system? At its core, transparency is conceived as the ability to see through or understand the inner workings and logic behind AI decisions. It's the antithesis of the so-called "black box", wherein decisions are made by the AI, but the reasons behind those decisions are inscrutable. While the idea seems straightforward, the definition and practical implications of transparency become multifaceted when looked at through different lenses. From a technical perspective, transparency often translates to understanding the algorithmic operations. It's about opening the black box and deciphering the mathematical and computational logic behind every decision. This viewpoint places importance on the intelligibility of the model's architecture, the nature of data training, and the dynamics of decision-making within the AI system. By contrast, the STS perspective on transparency encompasses a broader tableau. It's not merely about algorithmic clarity but extends to considerations of societal impact, ethical ramifications, and the accountability structure surrounding AI. For instance, even if a deep learning model is made technically transparent if the societal implications or potential biases are not laid bare, the model may still be deemed "opaque" from an STS standpoint. This divergence in understanding highlights the intricate nature of defining transparency and emphasizes the necessity to bridge the gap between technical and societal expectations. AI transparency, as conceptualized in a considerable portion of technical literature, is often viewed as a singular, cohesive construct with a unified definition and solution. However, an examination through the lens of STS offers a more nuanced perspective, breaking down the monolithic idea of transparency into diverse forms and types. The overarching tendency within some technical communities to pigeonhole transparency into a one-dimensional construct stems from a primary focus on algorithmic complexity. Such an approach not only simplifies the vast terrain of AI transparency but also perpetuates the fallacy that a singular technical solution exists that can universally ensure transparency. Contrarily, insights from STS scholars underscore that AI transparency is not confined to mere algorithmic intricacies. The multifaceted nature of transparency spans across various facets of AI, encompassing design, development, and application. The issue, as STS studies highlight, isn't strictly technical. Therefore, banking solely on technical solutions in the realm of XAI to wholly unravel the intricacies of AI transparency is, in many ways, a reflection of technochauvinistic tendencies. Such singular interpretations, though prevalent in some technical circles, risk oversimplifying a multifaceted challenge. By integrating insights from

STS scholarship, there emerges a clearer path towards adequately addressing the various types of opacity inherent in AI, facilitating a more holistic understanding and solution-oriented approach to AI transparency.

The dichotomy of transparency manifests distinctly in the two case studies highlighted previously. For instance, the music recommender system, while potentially transparent to its creators, may be opaque to the wider public. This could be a strategic decision by the platform, aligning with what Burrell terms "Opacity as intentional corporate or state secrecy" in Burrell (2016). In contrast, the open-source medical assistant, which is ostensibly transparent in its design, may yet be impenetrable to laypeople. This opacity could arise from its intrinsic structural complexity or from the specialized expertise needed to decipher its code, resonating with Burrell's classifications of "Opacity as the way algorithms operate at the scale of application" and "Opacity as technical illiteracy", respectively. This underscores the fact that even when systems have elements of transparency, the nature and implications of that transparency can differ. Without a nuanced understanding of these distinctions, one risks oversimplifying the multifaceted challenges and solutions associated with AI transparency and opacity.

5.1.2 When Every Answer Becomes an 'Explanation': The Pitfalls of Ambiguous Terminology

Having discussed the intricacies and ambiguities associated with the term 'transparency' in the preceding section, it becomes evident that the challenge of unclear definitions isn't isolated to just one facet of XAI. Another term that suffers from a similar lack of clarity, and is arguably at the very heart of the discipline, is 'explanation'. Just as the nebulous boundaries of transparency can lead to misconceptions, the ill-defined nature of explanation compounds the challenges faced by both developers and users of AI systems. Without a universally accepted definition, almost any form of supplementary information about an AI system's decision-making process can be branded as an 'explanation', regardless of its quality or comprehensibility. This indiscriminate labeling can have several adverse consequences. For one, it dilutes the genuine efforts being put into crafting meaningful explanations that are both accessible and informative. An explanation that doesn't elucidate the AI's reasoning, or is indecipherable to the end-user, effectively defeats its purpose. True explanations should serve to bridge the gap between AI processes and human understanding. They ought to provide insight, engender trust, and empower users to make informed decisions based on the AI's output. Moreover, by allowing any and all forms of information to qualify as explanations, we risk creating a landscape where the objective of achieving genuine interpretability and accountability is overshadowed by the mere act of providing data, regardless of its clarity or relevance. It's essential to recognize that the objective of explanations in XAI isn't just to provide more information but to deliver the right kind of information that aligns with the needs and comprehension levels of its audience. While efforts have been made to hone in on a more precise understanding of what constitutes an explanation, drawing insights from disciplines like philosophy, psychology, and cognitive science—as evidenced in works like Miller (2019)—the journey towards a unified definition remains ongoing. Such interdisciplinary endeavors provide a valuable foundation upon which a more structured and holistic field of XAI can be constructed. However, significant strides are still needed to fully encapsulate the depth and breadth of the term. In the upcoming sections of this chapter, we will delve deeper into the objectives of an explanation, considering its multifaceted nature and its connection to its intended audience. By dissecting the layers of interpretation, we aim to shed light on the nuanced role of explanations within XAI, emphasizing their pivotal significance in ensuring accountability, trust, and effective human-AI collaboration.

5.1.3 Terminology Inconsistency

A recurring challenge encountered in the literature, especially within the confines of technical discourse, is the inconsistent and sometimes interchangeable use of terms like transparency and interpretability. Each of these terms, while closely related to the other, encapsulates distinct aspects of understanding AI systems, and their conflation can lead to misinterpretation and lack of clarity in discussions. Transparency, as defined by scholars like Burrell (2016), pertains to the openness and accessibility of AI's processes. It's a prerequisite to understanding the internal workings of an AI system and ensures that stakeholders can see its mechanism and data operations without obfuscation. In essence, transparency addresses the "what" and "how" of AI operations. Interpretability, as outlined by Doshi-Velez and Kim (2017), and Ribeiro et al. (2016b) is about translating the complexities of an AI decision-making process into human-understandable terms. It doesn't necessarily delve deep into the system's internal intricacies but instead focuses on presenting outputs and decisions in a manner that users can comprehend. The goal is to answer the "why" behind an AI's decision, making it relatable to human stakeholders. The term "*explainability*" is also frequently used interchangeably with "*interpretability*". While these concepts share a thematic alignment, they possess nuanced differences. On the one hand, interpretability, as characterized in some works (and also discussed above), primarily denotes a human's capacity to understand the decision-making mechanism of an AI system. In this context, an interpretable model is one whose decisions can be readily comprehended by an individual, without necessarily requiring any ancillary information. Explainability, on the other hand, usually necessitates the presence of a third-party mechanism or tool that offers supplementary information to elucidate the decisions made by the AI system. It is through the conduit of explainability that interpretability is often achieved, although other avenues might also lead to interpretability without explicitly employing explainability. In Miller (2019), Miller delves into this nuanced distinction. Drawing upon Lipton's assertion in Lipton (2018), the paper aligns with the idea that explanations offer post-hoc interpretability. Further, referencing Biran and Cotton's definition from Biran and Cotton (2017), the paper defines the interpretability of a model as "the degree to which an observer can understand the cause of a decision." It underscores that while explanations serve as a vehicle to achieve understanding, other modes, like inherently understandable decisions or introspection, also exist. In their exploration of the terminological nuances, Miller ultimately aligns the concepts of 'interpretability' and 'explainability', a position that finds resonance among various scholars in the field. This thesis also subscribes to this alignment, primarily because the demarcation between interpretability and explainability, while notable, is not as critically consequential as the distinction between these terms and 'transparency'. Erroneously using 'interpretability' or 'explainability' as synonyms for 'transparency' could introduce ambiguities, potentially undermining the precision and clarity essential for the robust development of the XAI domain. When these terms are used interchangeably, it can skew the precision and clarity of discussions on AI. For instance, a model might be very transparent, providing full access to its code and data, but its decisions might remain arcane due to its complexity, making it less interpretable. To illustrate this event further, let's consider the medical assistant outlined in the second case study. This system, despite offering access to its code, could still pose challenges in terms of interpretability for the medical professionals who interact with it. This lack of interpretability might arise due to the structural complexity of the system or the specialized expertise necessary to comprehend its underlying code. This scenario aligns closely with the various forms of transparency discussed earlier, shedding light on the distinctions between transparency and interpretability. Moreover, it underscores the interdependence between these two concepts. While transparency can pave the way for interpretability, it doesn't guarantee it, highlighting the need to delve into the nuanced differences and connections between these aspects when addressing AI's transparency and opacity. Conversely,, a decision tree model used for loan approvals might offer clear, step-by-step decision criteria that a loan officer can easily follow and explain to applicants, rendering it highly interpretable. Yet, if the proprietary algorithm behind that model is kept hidden by the financial institution, it lacks transparency. Also, as previously highlighted, the music recommender system in our case study might be underpinned by a straightforward rule-based model, inherently rendering it interpretable. However, factors such as corporate discretion and proprietary protection could introduce opacity, obscuring its inner workings from external scrutiny. Recognizing and consistently adhering to these distinctions is crucial for fostering a clear, productive dialogue about AI's capabilities and challenges.

The inconsistency in terminology and ambiguous definitions of central terms in XAI has far-reaching implications. One of the most pressing concerns is that it opens a gateway for stakeholders to maintain the opacity of AI systems under the guise of transparency and interpretability. For instance, by merely offering some form of information, stakeholders might contend that they have provided an 'explanation', even when such information does little to clarify the AI's decision-making process for end-users. Similarly, the release of source code, while a gesture towards transparency, does not necessarily translate to interpretability for a majority who lack the technical expertise to decipher it, or because the source code alone does not provide the necessary information to interpret the system's decisions. This ambiguity can pose significant legal challenges. Regulations such as the Right to Explanation under the General Data Protection Regulation (GDPR) in the European Union, and the Algorithmic Accountability Act in the U.S., mandate the explainability of automated decisions. Misrepresentations of transparency and interpretability not only risk regulatory non-compliance but also endanger the trust and protection of the general public against the unchecked implications of opaque AI systems.

5.2 The Ripple Effect: How Unclear Terms Obscure XAI Aims and Users

Building on the preceding discussions, it becomes evident that the lack of clear definitions and the inconsistent use of pivotal terminology in XAI are not merely academic quibbles; they have tangible implications for the field's direction and efficacy. The ambiguity surrounding fundamental terms such as 'transparency' and 'explanation' is not just a matter of semantics. It poses substantial challenges, muddying the waters of understanding and making it difficult to establish clear objectives and to identify specific end-users for XAI systems. As mentioned in Lipton (2018) "Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable. Despite this ambiguity, many papers proclaim interpretability axiomatically, absent further explanation." When terms are fluidly defined or interchanged without precision, it leads to a scenario where there's no universal benchmark for what constitutes a valid explanation or a transparent system. In such an environment, almost any supplementary information about an AI system could be heralded as an explanation, regardless of its utility or comprehensibility to end-users. Without a clear target, the design, development, and evaluation of XAI systems become nebulous tasks. The inherent risk is that without specific objectives, XAI can veer off its foundational purpose: to make AI more comprehensible and accountable to its human stakeholders.

5.2.1 Objectives Lost in Translation

The adage, "If you don't know where you are going, any road will take you there," seems particularly apt when examining the trajectory of XAI in the face of inconsistent terminology and unclear definitions. With the fundamental terms of the field like 'transparency', 'explainability', and 'interpretability' being used interchangeably or without rigorous definition, the ripple effects are profound, leading to a landscape where the goals and objectives of XAI become increasingly hazy. How can we define the purpose of a system aimed at enhancing an AI model's transparency, if we haven't defined transparency? One immediate consequence of this haziness is the proliferation of XAI systems with widely differing aims, often dictated more by convenience or current technological capabilities than by a systematic understanding of what users truly need. Without a standardized vocabulary and shared understanding, there's a tendency to build systems that provide some manner of explanation, but without ensuring that such explanations cater to the requirements of specific user groups or real-world scenarios. It's akin to constructing a building without a blueprint, where every floor might have a different layout, not necessarily aligned with the inhabitants' needs. Additionally, the absence of clear objectives makes it challenging to measure the success or efficacy of XAI systems. When objectives are diverse or ill-defined, the metrics for evaluation become similarly nebulous. Consequently, the benchmarks for what constitutes a "good" or "effective" explanation in AI are varied and often incommensurate, resulting in evaluations that might prioritize technical accuracy over comprehensibility or vice versa. In essence, the lack of terminological precision and clear definitions has set off a cascade of challenges, with the objectives of XAI becoming a moving target, eluding a universal consensus and making it difficult to ascertain the field's progress and impact. XAI research is replete with papers asserting that their methodologies amplify trust, accountability, transparency, fairness, trustworthiness, and more. A closer inspection, however, reveals that many simply offer some degree of information about an opaque AI model, swiftly classifying it as an explanation. This, in turn, is perceived to automatically achieve a wide array of desired attributes. This issue is symptomatic of the entangled relationship between ambiguous terminology and nebulous objectives in the field. The absence of well-defined objectives facilitates a broad interpretation of what qualifies as an explanation, further bolstered by the claim that such explanations inherently attain the desired characteristics. While it's uncontestable that explanations can bolster user trust in an AI system and enhance its accountability, trustworthiness, and transparency, it's a misconception to believe that any explanation would universally accomplish these goals. It's essential to delineate the specific reasons for explaining an AI model, tailoring the design, implementation, and evaluation of the explainer in line with the set objectives. Additionally, XAI objectives aren't always in harmony with one another; in fact they can sometimes be at odds with each other Menis Mastromichalakis et al. (2024). This underscores the necessity to clearly articulate the primary goal and strategize accordingly. A case in point to illustrate the potential conflicts among XAI objectives would be the interplay between transparency and trust. Consider an AI system deployed in the banking sector for approving or rejecting loan applications. For the sake of transparency, let's say the AI system provides highly detailed explanations for every decision it makes. It might list out numerous factors, including small nuances in credit scores, intricate patterns in spending behavior, and deep statistical relationships, among others. While this depth of explanation might satisfy the objective of transparency, it could inadvertently undermine trust in the system. For a loan officer or a customer, such intricate details can be overwhelming, confusing, or even intimidating. On the one hand, an overabundance of information might lead them to mistrust the AI, not because it's wrong, but because its reasoning is too complex to easily follow or understand. On the other hand, if the AI system provided a simpler, high-level explanation to improve trust (e.g., "The loan was rejected due to a low credit score and recent large expenditures"), it might sacrifice some aspects of transparency. In this case, the underlying intricacies of the decision aren't fully disclosed. This example highlights the tension between transparency and trust. While achieving maximum transparency, we might inadvertently reduce trust, and while trying to foster trust through simplicity, we might compromise on full transparency. Similarly, while transparency is a desired characteristic of the medical assistant in our case study, especially for ensuring trust and informed decisions by the doctor, an excessively transparent explanation might not be ideal. Transparency, in this context, means providing a comprehensive account of why a specific treatment was suggested for the patient. However, the challenge arises when an explanation attempts to maximize transparency by providing all possible reasons leading to the decision, resulting in an overwhelming amount of information. If the doctor asks the medical assistant, "Why did you suggest this specific treatment for the patient?" and the goal of the assistant is to maximize transparency and list all the factors contributing to its decision, the explanation could include:

1. The patient is a human.
2. The patient's recorded height is within the average adult height range.
3. The patient has a heart rate within the range of 60-100 bpm.
4. The model hyperparameters were set to values derived from a previous grid search, including a learning rate of 0.001 and a dropout rate of 0.5.
5. Our embeddings were pre-trained on a corpus of 100,000 clinical reports.
6. Treatment A is recommended based on the activation patterns in layers 15 to 20 of the model.
7. The model achieved a ROC-AUC of 0.97 during training on a similar patient cohort.
8. The patient's blood sample shows elevated levels of marker X.
9. Genetic information indicates a mutation in gene Y.

10. Research suggests treatment A is effective for elevated levels of marker X.
11. The patient's shoe size is within the average range for adults.
12. Treatment A also counteracts the side effects of mutation in gene Y.
13. MRI scans showed abnormalities in the left lobe that treatment A targets.
14. Blood pressure readings indicate stability suitable for treatment A.
15. Historical data shows 60% of patients in the same age group respond well to this treatment.
16. The patient's last blood test indicated good kidney function, a prerequisite for this treatment.
17. Treatment A is available and in stock in the hospital's pharmacy.

This explanation, while it might be factually correct, mixes meaningful points with trivial and highly technical ones. On the one hand, instead of increasing the practitioner's trust, it could have the opposite effect due to the irrelevant information and the technical details that may not be interpretable by the doctor. On the other hand, if the objective of the explanation were to inform the doctor properly with relevant information to support the suggested treatment, it would focus on conveying the key details that are comprehensible to the doctor. For example, rules 8-10 and 12-17 might be the most relevant ones, presented in terminology familiar to the doctor. In practice, conveying only rules 8-10 and 12-13 might be sufficient to persuade the doctor of the validity of the treatment suggestion. However, it's important to recognize that fixating solely on a singular objective, such as trust, can carry significant risks. An undue emphasis on trust can inadvertently lead to deleterious consequences, particularly when it enables false or misleading explanations. Research studies, like the one conducted by Langer in Langer et al. (1978), have illuminated the fact that the mere presence of a justification or explanation can bolster trust, especially in non-critical scenarios. This phenomenon implies that if the sole aim of an explainer is to maximize trust, it could resort to furnishing explanations that are not only misleading but downright false, all in the pursuit of aligning with the end-user's intuition and comprehension. Consequently, users might develop a misplaced trust in the AI system, believing it to operate in a manner vastly different from reality. To elucidate this point, consider the scenario of our medical assistant in the case study. If unbeknownst to the user, the AI harbors an undesirable gender bias and bases its diagnostic decisions on the gender of the patient, a truthful explanation might reveal this bias. While it's ethically correct to expose such a flaw and it would be an accurate explanation, the consequences for user trust could be catastrophic, leading to a justified erosion of trust in the AI system. Paradoxically, in the pursuit of enhancing trust, the explainer would find itself in the morally murky territory of lying to users. Hence, it's evident that a multifaceted approach to explanation objectives is crucial. Beyond trust, considerations of transparency, relevance, accuracy, and fairness must also find their place. The challenge lies in balancing these objectives to ensure that explanations do not only increase the user's trust, but are also reliable, informative, and equitable.

5.2.2 Ignoring the End-Users

In the realm of XAI, one of the predominant oversights appears to be the neglect of the end-user. As developers and researchers work diligently to devise sophisticated AI models and the attendant explanations, there is an implicit assumption that greater transparency will naturally lead to better understanding. However, without tailoring explanations to the specific needs, backgrounds, and expertise of the intended audience, such efforts may yield limited results. In fact, they can inadvertently create a chasm between the model's intent and the user's comprehension. For instance, an explanation suitable for a data scientist, steeped in the intricacies of neural networks and machine learning, would likely be vastly different from one meant for a frontline healthcare professional using an AI-powered diagnostic tool. The former might crave mathematical precision and algorithmic details (e.g. rules 4-7 in the above example), while the latter seeks actionable insights in clear, jargon-free language (e.g. the rules considered relevant for the practitioner in the above example). Overloading the latter with technical nuances or treating the former with oversimplified details can both lead to a breakdown in trust and potential misuse of the AI tool. The irony here is clear: an endeavor to elucidate can further obfuscate if the end-user is not kept front and center. If a physician cannot intuitively understand why an AI recommends a particular treatment, they may dismiss the suggestion, potentially overlooking valuable insights. Ignoring the end-user's perspective doesn't just make explanations ineffectual; it can jeopardize the application of AI in critical sectors.

In light of the foregoing discussion, it becomes evident that an explanation's efficacy hinges on two pivotal elements: the provision of *relevant* information and the articulation of that information in a manner attuned to the end-user's understanding. Drawing from the illustrative case of the AI medical assistant, we discern distinct informational needs and comprehension levels for varied stakeholders. A practitioner, for instance, may grapple with the dense technicalities inherent in an explanation crafted for a data scientist, finding it both unintelligible and irrelevant. Conversely, the data scientist might find themselves adrift in a sea of arcane medical terminology. Moreover, certain explanations proffered in subsymbolic formats, like raw pixels or sound waves, might resonate with specific users while being enigmatic to others. This dichotomy not only underscores the significance of the explanation's format but also the necessity of its relevance. The multifaceted nature of explanations, wherein multiple causes can coexist, further underscores the imperative of relevance. To elucidate, Miller (2019) references a compelling example from Hanson (1965). In the aftermath of a fatal car accident, *"consider how the cause of death might have been set out by the physician as 'multiple haemorrhage', by the barrister as 'negligence on the part of the driver', by the carriage-builder as 'a defect in the brakelock construction', by a civic planner as 'the presence of tall shrubbery at that turning'. None is more true than any of the others, but the particular context of the question makes some explanations more relevant than others."* It's pivotal to emphasize that interpretability isn't an isolated attribute of an AI system; it's inherently relational, contingent upon the user in question. An AI model's interpretability is contingent upon its interplay with a specific user or user group. Miller (2019) accentuates this user-centric ethos, positioning XAI at the intersection of Social Sciences, AI, and Human-Computer Interaction, as elucidated in Figure 4.3.

5.3 What makes a "good" explanation? The challenges of XAI evaluation

Tightly interwoven with the aforementioned discussions, particularly the ambivalence surrounding the clear definition of objectives, lies another issue predominantly raised by the technical community yet scarcely touched upon by STS scholars: the evaluation of explanations. Despite the proliferation of methods and approaches in the realm of explainability, the yardstick for assessing their merit remains nebulous. The pertinent question arises: How does one quantify or qualify the efficacy of an explanation? What markers or metrics can be employed to discern a "good" explanation from a suboptimal one? Miller, in Miller (2019), observes that a majority of work in explainable AI predominantly hinges on researchers' subjective intuition in determining the quality of an explanation. There is, however, a rich repository of research in philosophy, psychology, and cognitive science that delves deep into the mechanisms of how individuals define, generate, select, evaluate, and communicate explanations. This body of work accentuates the cognitive biases and societal expectations inherent in the process of explanation. Just as we refrain from unconditionally accepting AI systems, it's essential to exercise the same skepticism towards the mechanisms elucidating their opaque functionalities. As highlighted in our recent work Filandrianos et al. (2023), the absence of a standardized framework for evaluating explanations frequently culminates in comparing disparate explanation methods, akin to juxtaposing apples with oranges. Although there exist some metrics to gauge the quality of explanations, they predominantly target specific types of explanations, emphasizing the fidelity of the explainer to the original AI system. However, as we've iterated earlier, the domain of explanation transcends mere technicality. The quintessential objective of an explanation is to enlighten an end-user about the intricacies of a given AI system. This outcome, unfortunately, remains largely unevaluated, with no consistent, universally accepted evaluation framework in place to assess explanations' impact on end-users. A comprehensive evaluation might entail user-centric surveys, deriving insights from disciplines spanning the social and cognitive sciences, psychology, and even philosophy. While there have been technically-driven efforts to assess explanations, they undeniably bring value to the table. For instance, the framework introduced in Pruthi et al. (2022), which quantifies the value of explanations through accuracy gains on a simulated model, stands as a testament to the ingenuity of technical solutions in this domain. Such approaches can effectively serve as proxies, providing valuable insight into the evaluative aspects of XAI. However, despite their inherent worth, they cannot be a substitute for direct user evaluations, which cater to the human element inherent in the process of explanation. Notably, there has been an emerging trend in the literature that underscores the significance of the human-centric perspective on explanations. The study by Ruth Byrne (Byrne, 2023), for instance, discusses the issue of what is a good explanation in XAI, using evidence from human explanatory reasoning. This work converges with the approach undertaken by Miller (2019) and discussed in chapter 4. Both studies, in their unique ways, pivot towards understanding explanations from the vantage point of human cognition and reasoning. They delve deep into the intricate mechanics of how individuals perceive, process, and assimilate explanations in their cognitive frameworks. This alignment of approaches underscores the growing recognition of the indispensability of integrating insights from cognitive and social sciences to truly capture the essence of what constitutes a "good" explanation. As the discourse on XAI evaluation evolves, these interdisciplinary bridges become even more crucial, emphasizing the holistic understanding of explanation from both technical and humanistic perspectives.

5.4 Beyond Technicalities: XAI's Interdisciplinary Imperative and the Politics of XAI

In the pursuit of unraveling the intricacies of explainable AI, it becomes apparent that XAI is not solely a technical domain but stands at the intersection of multiple disciplines. One particularly illustrative perspective comes from Miller's representation, which postulates XAI as the confluence of AI, Human-Computer Interaction (HCI), and Social Sciences (see figure 4.3 from Miller (2019)). This intersection offers a rich tapestry of insights, drawing from the algorithmic depth of AI, the user-centric focus of HCI, and the behavioral and cognitive dimensions encapsulated within the Social Sciences. The intertwining of these disciplines in the realm of XAI signifies the multi-faceted nature of the challenges and solutions inherent in the field. From AI, we derive the foundational understanding of models and algorithms, understanding their operational intricacies and potential for enhancement. HCI brings to the table the principles of user experience and interface design, emphasizing the importance of tailoring explanations to the specific needs, contexts, and comprehension levels of different user groups. Simultaneously, the Social Sciences furnish us with profound insights into human cognition, decision-making, and trust dynamics, pivotal for ensuring that AI explanations resonate genuinely with the human psyche. The multidisciplinary approach to XAI is not just about integrating technical, cognitive, and human-computer interaction perspectives. It's equally important to recognize the social and political dimensions that shape, and are shaped by AI systems. Drawing inspiration from Winner's seminal work "Do Artifacts Have Politics?" (Winner, 1980), we can delve deeper into the intertwined nature of technology and societal structures. Winner compellingly challenges the notion that technologies are neutral tools, positing instead that they can inherently embody specific forms of power and authority. Within the XAI context, this implies that the very nature of interpretability transcends pure technical constraints. While we've established that interpretability is deeply linked to the end-user and their environment, there's an added layer of complexity brought about by the socio-political milieu within which an AI system is conceived, developed, and deployed. Such systems aren't devoid of the politics and values of their creators and the institutions that birthed them. These systems mirror the goals, biases, and aspirations of their designers and, at times, the broader societal structures they stem from. Thus, when discussing AI interpretability, it's not merely a matter of determining if a technology is functional. We must delve deeper, asking how an AI system, or its explanation, shapes and is influenced by societal structures, norms, and values. Decisions in design, implementation, and deployment can either perpetuate existing power dynamics or, ideally, democratize access and understanding. In aligning with Winner's perspective, we emphasize that AI systems and their explanations aren't apolitical or neutral entities. They carry with them the weight of societal structures, power dynamics, and decision-making frameworks. As we advocate for the development and deployment of more interpretable AI systems, we must remain cognizant of these intertwined political and social dimensions. Engaging in this deeper interrogation ensures a holistic and comprehensive approach to XAI, reinforcing the idea that the quest for interpretability is as much (if not more) a societal endeavor as it is a technical one. Tackling XAI merely from a technical angle can certainly facilitate advancements in algorithmic clarity, but doing so risks oversimplifying the challenges and sidelining the diverse needs of end-users. Further, by ignoring the political and societal nuances of AI systems, we may inadvertently perpetuate existing power dynamics and biases. A holistic, interdisciplinary approach to XAI ensures that we are not just focusing on mathematical precision and rigor but also accounting for the complex cognitive, emotional, and societal intricacies that shape human interaction with technology. To illustrate the importance of approaching AI systems as an artifact with politics, let's consider our first case study of the music recommender system. A system like Spotify's music recommender isn't just a tool that understands music preferences; it's also influenced by:

1. Music Industry Relations: It might recommend tracks from labels with which Spotify has better financial arrangements or from artists who are currently being heavily promoted.

2. Cultural Biases: If predominantly designed by engineers from Western countries, it might have an implicit bias towards Western musical tastes, unintentionally sidelining artists from other regions.

3. User Base Influence: If the majority of Spotify users are from a certain age group, the AI is predominantly trained on their preferences, possibly skewing recommendations for older or younger users.

An explanation from Spotify that acknowledges these influences would paint a picture of why a certain song is suggested – it's not just the user's preference, but a combination of business relations, cultural biases, and predominant user base trends. When viewed in isolation, the system might simply state that it recommended a song because it matches the user's listening habits and shares characteristics with songs the user has enjoyed. With the politics and socio-economic influences taken into account, the system's explanation might expand: "This song matches your listening habits and shares characteristics with songs you've enjoyed. It should be highlighted, though, that the track comes from a label with which we have a current promotional agreement, and our algorithm has a tendency to favor Western musical tastes due to its primary design influences." In the same rationale, consider the second case study, of the medical assistant for oncologists. When examining the medical assistant from an STS lens that includes its political and environmental origins, one recognizes it's not just a machine-learning model designed to assist oncologists. It embodies a multitude of decisions made at various levels:

1. Healthcare Policies: Developed possibly in an environment where early cancer detection is a high priority, the assistant's algorithms may inadvertently prioritize certain types of diagnoses over others due to governmental health directives or funding incentives.

2. Medical Training: The medical knowledge it was trained on comes from institutions with their own biases – some treatments might be preferred over others, not necessarily because they're universally better, but due to regional or institutional preferences.

3. Data Collection: The data the model was trained on can carry biases. If predominantly trained on data from affluent patients, the model might not perform as well when confronted with data from patients with different socio-economic backgrounds.

An explanation from this AI system that encapsulates these factors will not only provide diagnostic reasoning but also incorporate why a certain treatment is favored, the biases the model might have, and the societal structures that influenced its design. Be that as it may, if one were to simply view the assistant as a standalone system, the assistant might explain that it's suggesting a treatment because a particular mutation was found in a gene or due to certain patterns recognized in an MRI scan. However, when factoring in the politics and socio-economic structures that influenced its design, the explanation might be extended to mention: "This treatment is suggested based on gene mutation and MRI patterns. It's worth noting that the treatment is among those heavily prioritized in recent years due to current healthcare policies, and our database has an over-representation of affluent patient data which might influence the recommendation." It's evident that the choices and behaviors of AI systems, which might initially seem grounded purely in technical logic or appear as benign technical flaws, gain a clearer dimension when viewed through the lens of their socio-political context. To fully understand and critically assess these systems, one must consider this broader backdrop. By folding in the politics and broader influences of each artifact into their explanations, users are provided a more holistic view, which gives them better insight into the potential biases and external factors influencing AI decisions. This perspective is crucial for moving towards AI implementations that are not only fair and inclusive but truly trustworthy. As we continue our journey in the domain of XAI, it's paramount to acknowledge its multidimensional nature. This recognition acts as our compass, directing us towards the creation of AI systems that are not only transparent and precise but also deeply resonant, inclusive, and mindful of the broader socio-political landscape they operate within.

Chapter 6

Concluding Thoughts and The Road Ahead

Throughout this thesis, we embarked on an exploratory journey into the realm of Explainable AI, viewed through the interdisciplinary lens of Science and Technology Studies (STS). This venture was foundational in reshaping our understanding of XAI, nudging it beyond the confines of pure technicality and situating it firmly within broader societal contexts. From our comprehensive review of STS literature, we recognized the innate complexities and multifaceted issues surrounding technological developments, particularly AI. These insights were then juxtaposed with in-depth analyses from technical literature on XAI. This methodology allowed us to uncover critical intersections and dissonances, guiding our quest to align technological advancements with societal implications. Through the marriage of these two domains, we distilled key insights, highlighting gaps, synergies, and most importantly, the value of an interdisciplinary approach to AI's explainability challenge. To bridge the gap between theory and practice, we employed case studies and empirical analyses. These not only anchored our theoretical findings into tangible scenarios but also illuminated the broader impacts and practicalities of integrating STS perspectives into XAI advancements. Said integration underlined the necessity for AI developments to be not just technically sound, but also socially responsible and attuned to human intricacies. Looking ahead, the rich tapestry of insights woven throughout this thesis paves the way for numerous avenues of future research. There remains vast potential in further unpacking the intersections of STS and XAI, particularly as AI technologies evolve and are embedded even deeper into societal frameworks. Future explorations could delve deeper into user-centric evaluations, understanding biases inherent in AI systems, or exploring how different cultures and societies interpret and value explanations. In essence, as AI continues to redefine our world, ensuring its explainability and interpretability remains an interdisciplinary challenge, necessitating collaborative, informed, and continuous efforts. By synthesizing the wisdom from both STS and technical perspectives, this thesis aims to serve as a beacon for researchers, technologists, and policymakers alike, urging them towards the co-creation of a future where AI is not just advanced, but also aligned with the societal fabric it is intertwined with.

Bibliography

Gloria Andrada, Robert W Clowes, and Paul R Smart. Varieties of transparency: Exploring agency within ai systems. *AI & society*, 38(4):1321–1331, 2023.

Brian Ball and Alexandros Koliousis. Training philosopher engineers for better ai. *AI & SOCIETY*, 38(2):861–868, 2023.

David Beer. The social power of algorithms. In *The Social Power of Algorithms*, pages 1–13. Routledge, 2019.

Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017.

Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, pages 1–60, 2023.

Meredith Broussard. *Artificial unintelligence: How computers misunderstand the world.* mit Press, 2018.

Meredith Broussard. *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech.* MIT Press, 2023.

Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big data & society*, 3(1):2053951715622512, 2016.

Ruth M. J. Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6276–6282. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/876. URL https://doi.org/10.24963/ijcai.2019/876.

Ruth MJ Byrne. Good explanations in explainable artificial intelligence (xai): Evidence from human explanatory reasoning. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. International Joint Conferences on Artificial Intelligence Organization, 2023.

Manuel Carabantes. Black-box artificial intelligence: an epistemological and critical analysis. *AI & society*, 35(2):309–317, 2020.

Nathan Colaner. Is explainable artificial intelligence intrinsically valuable? *AI & SOCIETY*, pages 1–8, 2022.

Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *CoRR*, abs/2006.11371, 2020. URL https://arxiv.org/abs/2006.11371.

Karl de Fine Licht and Jenny de Fine Licht. Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy. *AI & society*, 35:917–926, 2020.

Daniel C Dennett. *The intentional stance*. 1989.

Edmund Dervakos, Orfeas Menis-Mastromichalakis, Alexandros Chortaras, and Giorgos Stamou. Computing rule-based explanations of machine learning classifiers using knowledge graphs. *arXiv preprint arXiv:2202.03971*, 2022.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, 31, 2018.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

George Filandrianos, Edmund Dervakos, Orfeas Menis Mastromichalakis, Chrysoula Zerva, and Giorgos Stamou. Counterfactuals of counterfactuals: a back-translation-inspired approach to analyse counterfactual editors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9507–9525, 2023.

Alex A. Freitas. Comprehensible classification models: A position paper. *SIGKDD Explor. Newsl.*, 15(1):1–10, mar 2014. ISSN 1931-0145. doi: 10.1145/2594473.2594475. URL https://doi.org/10.1145/2594473.2594475.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Mario Günther and Atoosa Kasirzadeh. Algorithmic and human decision making: for a double standard of transparency. *AI & SOCIETY*, pages 1–7, 2022.

Thilo Hagendorff and Katharina Wezel. 15 challenges for ai: or what ai (currently) can't do. *AI & SOCIETY*, 35:355–365, 2020.

Orit Halpern, Robert Mitchell, and Bernard Geoghegan. The smartness mandate: Notes toward a critique. *Grey Room*, 68:106–129, 09 2017. doi: 10.1162/GREY_a_00221.

Norwood Russell Hanson. *Patterns of discovery: An inquiry into the conceptual foundations of science*. CUP Archive, 1965.

Clément Henin and Daniel Le Métayer. Beyond explainability: justifiability and contestability of algorithmic decision systems. *AI & SOCIETY*, pages 1–14, 2021.

Tomasz Hollanek. Ai transparency: a matter of reconciling design with critique. *AI & SOCIETY*, pages 1–9, 2020.

Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011. ISSN 0167-9236. doi: https://doi.org/10.1016/j.dss.2010.12.003. URL https://www.sciencedirect.com/science/article/pii/S0167923610002368.

Daniel Innerarity. Making the black box society transparent. *AI & SOCIETY*, pages 1–7, 2021.

Christian Katzenbach and Lena Ulbricht. Algorithmic governance. *Internet Policy Review*, 8(4):1–18, 2019.

Ellen J Langer, Arthur Blank, and Benzion Chanowitz. The mindlessness of ostensibly thoughtful action: The role of" placebic" information in interpersonal interaction. *Journal of personality and social psychology*, 36(6):635, 1978.

Emma E Levine and Maurice E Schweitzer. Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126:88–106, 2015.

Jason Liartis, Edmund Dervakos, Orfeas Menis-Mastromichalakis, Alexandros Chortaras, and Giorgos Stamou. Semantic queries explaining opaque machine learning classifiers. In *DAO-XAI*, 2021.

Jason Liartis, Edmund Dervakos, Orfeas Menis-Mastromichalakis, Alexandros Chortaras, and Giorgos Stamou. Searching for explanations of black-box classifiers in the space of semantic queries. *Semantic Web*, (Preprint):1–42, 2023.

Zachary C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, jun 2018. ISSN 1542-7730. doi: 10.1145/3236386.3241340. URL https://doi.org/10.1145/3236386.3241340.

Felix Tun Han Lo. The paradoxical transparency of opaque machine learning. *AI & SOCIETY*, pages 1–13, 2022.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Maria Lymperaiou, George Manoliadis, Orfeas Menis Mastromichalakis, Edmund G Dervakos, and Giorgos Stamou. Towards explainable evaluation of language models on the semantic similarity of visual concepts. *arXiv preprint arXiv:2209.03723*, 2022.

Orfeas Menis Mastromichalakis, Jason Liartis, and Giorgos Stamou. Beyond one-size-fits-all: Adapting counterfactual explanations to user objectives. *ACM CHI Workshop on Human-Centered Explainable AI (HCXAI)*, 2024.

Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2018.07.007. URL https://www.sciencedirect.com/science/article/pii/S0004370218305988.

Andres Morales-Forero, Samuel Bassetto, and Eric Coatanea. Toward safe ai. *AI & SOCIETY*, 38(2): 685–696, 2023.

Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. Accountability in artificial intelligence: what it is and how it works. *AI & SOCIETY*, pages 1–12, 2023.

Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

Guglielmo Papagni, Jesse de Pagter, Setareh Zafari, Michael Filzmoser, and Sabine T Koeszegi. Artificial agents' explainability to support trust: considerations on timing and context. *AI & SOCIETY*, 38(2):947–960, 2023.

Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. Meaningful explanations of black box ai decision systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9780–9784, Jul. 2019. doi: 10.1609/aaai.v33i01.33019780. URL https://ojs.aaai.org/index.php/AAAI/article/view/5050.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C Lipton, Graham Neubig, and William W Cohen. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375, 2022.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-3020. URL https://aclanthology.org/N16-3020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016b.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11491. URL https://ojs.aaai.org/index.php/AAAI/article/view/11491.

Amber Ross. Ai and the expert; a blueprint for the ethical use of opaque ai. *AI & SOCIETY*, pages 1–12, 2022.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 05 2019. doi: 10.1038/s42256-019-0048-x.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Donghee Shin, Joon Soo Lim, Norita Ahmad, and Mohammed Ibahrine. Understanding user sensemaking in fairness and transparency in algorithms: algorithmic sensemaking in over-the-top platform. *AI & SOCIETY*, pages 1–14, 2022.

Manolis Simos, Konstantinos Konstantis, Konstantinos Sakalis, and Aristotle Tympas. Ai can be analogous to steam power. *ICON*, 27(1):97–116, 2022.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Joel Walmsley. Artificial intelligence and the value of transparency. *AI & SOCIETY*, 36(2):585–595, 2021.

Adrian Weller. Challenges for transparency. 2017.

R Wiley. The evolution of communication: information and manipulation. *Animal behaviour*, 2: 156–189, 1983.

Langdon Winner. Do artifacts have politics? *Daedalus*, 109(1):121–136, 1980.

John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. Transparency in algorithmic and human decision-making: is there a double standard? *Philosophy & Technology*, 32:661–683, 2019.