LUNDS UNIVERSITET
Ekonomihögskolan

Institutionen för informatik

Transparency in the Black Box

The drug development industry's perspective on ethical concerns posed by Black Box AI, and how well it aligns with EU requirements

Kandidatuppsats 15 hp, kurs SYSK16 i Systemvetenskap.

Författare: Sofie Schnitzer
Mathewos Tiruwuha

Handledare: Christina Keller

Rättande lärare: Nam Aghaee
Benjamin Weaver

Transparency in the Black Box: The drug development industry’s perspective on ethical concerns posed by AI, and how well it aligns with EU requirements

SVENSK TITEL: Transparens i den Svarta Lådan: Läkemedelsbranschens perspektiv på etiska dilemman med AI och hur väl det stämmer överens med krav från EU.

FÖRFATTARE: Sofie Schnitzer, Mathewos Tiruwuha

UTGIVARE: Institutionen för informatik, Ekonomihögskolan, Lunds universitet

EXAMINATOR: Osama Mansour, Docent

FRAMLAGD: Maj, 2024

DOKUMENTTYP: Kandidatuppsats

ANTAL SIDOR: 101

NYCKELORD: Artificial Intelligence, Drug Development, Ethics, Explainability, EU Requirements

ABSTRACT:

The integration of Artificial Intelligence (AI) into drug development introduces significant opportunities and ethical challenges, particularly with the use of complex "Black Box" AI systems. This study investigates how the drug development industry addresses ethical concerns associated with AI, focusing on compliance with European Union (EU) regulations, including the Assessment List for Trustworthy Artificial Intelligence (ALTAI) and the Artificial Intelligence Act (AI Act). Through a literature review and qualitative interviews with industry professionals, this research explores how the industry manages the trade-off between harnessing AI's capabilities and meeting ethical standards for transparency. The findings highlight a strong industry emphasis on the importance of explainability in AI systems, which is crucial for building trust and adhering to regulatory frameworks. The thesis also examines the broader implications of these ethical issues for the future use of AI in drug development, proposing a framework that promotes both innovation and ethical compliance. The gathered perspectives largely reflect a proactive approach by the industry to meet EU standards, underscoring a commitment to ethical AI deployment. This research enhances the understanding of the relationship between technological progress and ethical guidelines in drug development, emphasizing the importance of transparency and explainability in meeting industry and regulatory expectations.

SAMMANFATTNING:

Integrationen av Artificiell Intelligens (AI) i läkemedelsutvecklingen medför betydande möjligheter och etiska utmaningar, särskilt med användning av komplexa "Black Box" AI-system. Denna studie undersöker hur läkemedelsindustrin hanterar etiska frågor kopplade till AI, med fokus på efterlevnad av EU:s regelverk, inklusive Assessment List for Trustworthy Artificial Intelligence (ALTAI) samt Artificial Intelligence Act (AI Act). Genom en litteraturöversikt och kvalitativa intervjuer med branschfolk utforskas hur industrin balanserar mellan att utnyttja AI:s kapacitet och att uppfylla etiska standarder för transparens. Resultaten betonar en stark industriell betoning på vikten av förklarbarhet i AI-system, vilket är avgörande för att bygga förtroende och följa regelverk. Uppsatsen undersöker också de bredare konsekvenserna av dessa etiska frågor för framtida användning av AI i läkemedelsutveckling, och föreslår ett ramverk som främjar både innovation och etisk efterlevnad. Insamlade perspektiv reflekterar i stor utsträckning en proaktiv inställning från industrin för att möta EU-standarder, vilket understryker ett åtagande för etiskt AI-implementering. Forskningen förbättrar förståelsen för sambandet mellan tekniska framsteg och etiska riktlinjer och perspektiv i läkemedelsutvecklingen, och betonar vikten av transparens och förklarbarhet.

Acknowledgements

We want to direct our gratitude to everyone who helped us with our research, your help is greatly appreciated. First, a big thanks to our supervisor, Christina Keller, for guiding us through this process with great advice and patience. We are also grateful to all the research participants for their time and insights, which were vital for our study. Lastly, thanks as well to those who helped connect us with experts in the drug development industry, your help was invaluable and deeply appreciated.

May 2024

Sofie Schnitzer & Mathewos Tiruwuha

1 Introduction 1
- 1.1 Background 1
- 1.2 Problem Statement 2
- 1.3 Research Question 3
- 1.4 Purpose 3
- 1.5 Delimitations 3
- 1.6 Motivation of Terminology 4
  - 1.6.1 Explainability 4
  - 1.6.2 Requirement Frameworks 4

2 Literature Review 5
- 2.1 e-Health 5
  - 2.1.1 Defining e-Health 5
- 2.2 Artificial Intelligence 5
  - 2.2.1 Defining Artificial Intelligence 5
  - 2.2.2 Ethical Artificial Intelligence 5
  - 2.2.3 Black Box Models 6
- 2.3 Drug development 6
  - 2.3.1 Original drug development processes 6
  - 2.3.2 Drug Development with Artificial Intelligence 10
- 2.4 Research Frameworks 13
  - 2.4.1 Explainable Artificial Intelligence (XAI) 13
  - 2.4.2 The Assessment List for Trustworthy Artificial Intelligence (ALTAI) by EU 15
  - 2.4.3 Artificial Intelligence Act (AI Act) by EU 16
- 2.5 Previous Research 18
  - 2.5.1 EU and the XAI Theory (Projects XAI and NL4XAI) 18
  - 2.5.2 European Data Protection Supervisor (EDPS); TechDispatch 18
  - 2.5.3 XAI in the AI Act 20
  - 2.5.4 Reflections by the European Medicine Agency on EU requirements for Artificial Intelligence in Medicine 20
  - 2.5.5 Guidelines for Artificial Intelligence in Swedish Healthcare 21
  - 2.5.6 Developers on Ethical Artificial Intelligence in Medicine Applications 22
  - 2.5.7 Healthcare Leaders on Artificial Intelligence in Healthcare 23
  - 2.5.8 Medical Technology Industry Perspective on AI Act 24
  - 2.5.9 Epistemological and Ethical Issues in the Trust of Black Box Algorithms in Medicine 24

3 Methodology 25
- 3.1 Choice of Method 25
- 3.2 Literature Search 26
- 3.3 Empirical Data Collection 27
  - 3.3.1 Setting 27
  - 3.3.2 Respondents 27
  - 3.3.3 Interview Guide 29

- 3.3.4 Ethical Considerations........30
- 3.4 Data Analysis........31
- 3.5 Validity and Reliability........35

4 Empirical Findings........37

- 4.1 Theme 1: Ethical AI and Transparency........37
- 4.2 Theme 2: Regulatory Alignment and Challenges........40
- 4.3 Theme 3: Stakeholder Perspectives and Industry Challenges........41
- 4.4 Theme 4: Balancing Innovation and Ethical Standards........42
- 4.5 Theme 5: Integration of AI in Clinical and Pharmaceutical Ethics........44

5 Discussion........47

- 5.1 Foundation of Theoretical and Regulatory Perspectives........47
  - 5.1.1 Requirements for Black Boxes and XAI common for ALTAI and AI Act........48
- 5.2 Perspectives of the Drug Development Industry........48
  - 5.2.1 Ethical AI and Transparency........48
  - 5.2.2 Regulatory Alignment and Challenges........49
  - 5.2.3 Stakeholder Perspectives and Industry Challenges........50
  - 5.2.4 Balancing Innovation and Ethical Standards........50
  - 5.2.5 Integration of AI in Clinical and Pharmaceutical Ethics........51

6 Conclusion and Recommendations........52

- 6.1 Conclusion........52
- 6.2 Limitations........53
- 6.3 Recommendations for Future Research........53

References........54

- Appendix A - Interview Guide........60
- Appendix B - Informed Consent Form........62
- Appendix C - Interview R1........64
- Appendix D - Interview R2........69
- Appendix E - Interview R3........74
- Appendix F - Interview R4........78
- Appendix G - Interview R5........85
- Appendix H - ChatGPT Prompt 1........92
- Appendix I - ChatGPT Output........93
- Appendix J - AI Contribution Statement........94

Figures

Figure 2.1: The traditional drug development process (Adapted from Tamimi & Ellis, 2009)....7

Figure 2.2: Visual representation of examples where AI can benefit the drug development lifecycle (Adapted from Tamimi & Ellis, 2009; modifications and comments added by the authors)....13

Figure 2.3: Visual representation of the concept of XAI (Adapted from McDermid et al., 2021)....14

Figure 2.4: AI Risk Level Framework (Adapted from European Commission, 2024)....17

Tables

Table 3.1: Respondent and Interview information…29

Table 3.2: Final table of themes…35

1 Introduction

In the following introduction we present the background to the research topic, formulate the problem statement as well as describe the purpose and delimitations of this study.

1.1 Background

The invention of chemical medications and the following rise of the pharmaceutical industry in the early 1900s were significant turning points in the history of drug development. Today, the innovative field of Artificial Intelligence (AI) is leading this industry through yet another dramatic change. AI has great potential for accelerating drug development, cutting costs, and eventually improving patient outcomes as we approach the dawn of this new age (Stahl et al., 2022). It is difficult not to envision the potential of AI in tackling diseases like cancer and Alzheimer's, which have long been dreaded and lethal diagnoses.

The notion of AI dates back to the 1950s, when researchers predicted that robots would eventually be able to mimic human intelligence (Arreche et al., 2024). Although there were early believers back then, the current reality of AI raises concern, especially in light of its potential surpassing human control and understanding (Arreche et al., 2024). The term ‘Black Box’ refers to AI models that function without offering an explanation for the reasoning behind their choices, which makes them incomprehensible to man (Anjomshoae, 2022; European Commission, 2020).

In the realm of drug development, AI's influence spans the entire spectrum, from the initial discovery of promising compounds to the post-approval surveillance of newly introduced medications (Chen et al., 2021). Traditionally, the process of bringing a novel drug to market entails a decade-long journey fraught with hurdles and expenses. However, the integration of AI promises to compress this timeline and mitigate costs, ushering in a new era of efficiency and efficacy (Chen et al., 2021).

Achieving a balance between harnessing the power of AI and maintaining an ethical approach remains a critical challenge in the ongoing development and deployment of AI systems, the pharmaceutical industry being no exception (Arreche et al., 2024). This complex navigation is currently being evaluated and regulated by governing powers, such as the European Union's efforts to provide guidance and a set of rules of conduct (European Commission, 2020; European Parliament and Council of the European Union, 2024).

1.2 Problem Statement

Recent advancements in AI have propelled the field of drug development into a new era, promising cost reduction and efficiency enhancement (Aliper et al., 2016; Stokes et al., 2020; Stahl et al., 2022). The integration of AI into drug development and healthcare has accelerated, with AI technologies becoming indispensable across various phases of the drug research and development process. However, alongside these advancements, the ethical challenges posed by AI, particularly in its opaque ‘Black box’ systems, are becoming increasingly apparent. Reliance on opaque AI systems raises significant ethical concerns. Machine learning and deep learning algorithms, characterized by their lack of transparency and interpretability, can introduce biases, errors, and discriminatory outcomes, undermining trust, and accountability.

One framework addressing ethical concerns posed by AI is the Assessment List for Trustworthy Artificial Intelligence (ALTAI), developed by the European Commission's High-Level Expert Group on AI (AI HLEG) (2020). ALTAI provides criteria for trustworthy AI. In addition, the European Parliament has introduced the AI Act, the world's first comprehensive regulatory framework on AI, aimed at promoting trustworthy AI usage (European Parliament and Council of the European Union, 2024). While both are designed for wider AI applications across various sectors, the principles of ALTAI and the AI Act are also pertinent to ethically integrating AI in drug development.

In the realm of drug development, where decisions directly impact human health and well-being, the opacity of AI systems poses significant concerns. While there is a consensus on the importance of ethical considerations in deploying AI models in healthcare, there are divergent opinions on the balance between ethics and innovation (European Commission, 2020; Weidener & Fischer, 2024). Despite the extensive literature on AI in drug development, an industry highly bound to ethics and heavily reliant on innovation, there is a notable absence of data concerning the drug development industry’s stance on the ethical use of AI and how it aligns with existing requirements.

Petersson et al.’s (2022) study on healthcare leaders' perspectives reveals that laws and policies have not kept pace with the rapid innovation and advancements of AI in healthcare, according to the industry. Participants in the study emphasize the necessity for research into specific areas of the healthcare sector, noting that different domains may face distinct AI-related challenges (Petersson et al., 2022). Similarly, Weidener and Fischer’s study (2024) stresses the need for further research into the ethical use of AI in medicine and the establishment of robust ethical and legal frameworks in this domain. This study aims to contribute by exploring the drug development industry’s viewpoint on Black Box AI and assessing its alignment with EU requirements, particularly those outlined in EU ALTAI and AI Act concerning AI explainability.

1.3 Research Question

In light of the increasing integration of AI into the pharmaceutical industry, particularly in drug development, we want to study how the drug development sector, currently utilizing AI, views the ethical issues posed by complex AI models. Also acknowledging the EU recommendations mandating requirements for the use of AI, notably for explainability and transparency, this research seeks to address the following question:

- *What is the drug development industry's perspective on the ethical concerns posed by Black Box models?*
- *How does this perspective correspond with EU requirements surrounding AI explainability?*

1.4 Purpose

The purpose of this research paper is to, based on our empirical data from qualitative interviews and additionally on literature reviews, highlight the drug development industry's perspective on ethical challenges with Black Box AI, and how their perspective aligns with the requirements regarding explainable AI that experts have developed for the EU Artificial Intelligence Act (AI Act) as well as the Assessment List for Trustworthy Artificial Intelligence (ALTAI). The purpose of this is to increase the understanding of ethical implications of AI models in drug development processes, especially from an industry perspective. Furthermore, the purpose is to see if the industry perspective aligns with current EU requirements regarding AI explainability.

1.5 Delimitations

The study is limited to the drug development industry utilizing AI and has focused the qualitative research on the perspectives of individuals active within the drug development sector who also possess understanding of how AI is used within the drug development industry. Additionally, this study aims to understand the industry's perspective on ethical dilemmas arising from using black box AI models and to compare these perspectives with current EU requirements specifically related to black box AI.

1.6 Motivation of Terminology

1.6.1 Explainability

In the AI Act, 'transparency' pertains to systems that ensure 'explainability' and 'traceability' of AI (European Parliament and Council of the European Union, 2024). The study focuses on Black Box models, which are specifically referred to in the ‘explainability’ section of the ALTAI framework (European Commission, 2020). Additionally, this study uses explainable AI (XAI) as the foundation for comprehensibility of AI systems. Based on this background, we will throughout this study primarily use, and focus on, the term ‘explainability’. However, 'explainability' is often used interchangeably with related terms in literature, therefore, similar terms such as ‘transparency’, ‘opaqueness’ and ‘interpretability’ will occur in this research.

1.6.2 Requirement Frameworks

The reasoning behind the terminological decision to use ‘requirements’ instead of ‘regulations’ when discussing the EU frameworks, is that the ALTAI is a requirement framework, and the AI Act is a not yet legislative regulatory framework (European Commission, 2020; European Parliament and Council of the European Union, 2024). Hence, we found the term requirements to be the most fitting when referring to both frameworks in this study.

2 Literature Review

The following literature review covers essential information about the area of research, theories used as well as previous research within the area.

2.1 e-Health

2.1.1 Defining e-Health

In the subject area of drug development and artificial intelligence, we are situated at the intersection of digitalization and health, an area referred to as e-Health. The World Health Organization (WHO) (n.d.) defines eHealth as including various health-related technologies such as telehealth, telemedicine, mobile health, electronic medical/health records, big data, wearables, and artificial intelligence. These innovations improve healthcare delivery by enabling remote services, enhancing data management, supporting real-time health monitoring, and advancing diagnostics and personalized treatments. eHealth is crucial for ensuring healthcare that is accessible, efficient, and effective (World Health Organization, n.d.).

2.2 Artificial Intelligence

2.2.1 Defining Artificial Intelligence

There is not a singular, widely accepted definition of artificial intelligence. However, we have decided to accept the following definition of AI for the purposes of this study. This definition of AI is aimed to close the definition gap (Gbadegeshin et al., 2021).

> *"AI is a 'system' not only a technology that can make use of data, learn by itself and act on its lessons to perform assigned task (s) effectively and efficiently in any environment. This system must be trained initially (either by humans or another system). It can automate a process, direct itself and continuously learn from its activities. It can also act appropriately, independently, and intelligently with little human input. It contains different forms of software and/or devices. It is created by humans. It is built on understanding of existing phenomena, and it acts wisely based on its understanding."* (Gbadegeshin et al.,2021, p.6)

2.2.2 Ethical Artificial Intelligence

AI will in this study be discussed in the context of ethics and ethical approaches to AI. We will use Leslie's (2019) definition to clarify the meaning of ethical AI.

"AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies." (Leslie, 2019, p.3)

2.2.3 Black Box Models

AI is gaining traction across various sectors including finance, transportation, and manufacturing, with healthcare being particularly noteworthy (European Data Protection Supervisor, 2023). Its ability to automate processes like analyzing large data sets and identifying patterns, together with its increased accessibility to the general public in recent years, are the main reasons for its rise in popularity (European Data Protection Supervisor, 2023).

Despite the growing adoption of AI, many systems operate opaquely, leaving providers, deployers, and those affected by AI unaware of their inner workings (European Data Protection Supervisor, 2023). An opaque system, as described by Besold and Uckelman (2018), is one in which everything occurring between input and output remains invisible to humans. Such systems make decisions and predictions without providing any insight into the reasoning behind their conclusions. This phenomenon is commonly known as Black Box systems (European Data Protection Supervisor, 2023; Besold & Uckelman, 2018).

The risk posed by the Black Box systems lies in their departure from explicit human programming; instead, AI technologies such as machine learning (ML) and deep learning (DL) systems rely on algorithms learned through self-training (European Data Protection Supervisor, 2023). To illustrate, consider the training phase, where AI models can uncover new relationships between specific input features (such as clinical symptoms) and generate predictions or decisions (such as medical diagnoses) using highly complex models with millions of interacting parameters (Peters, 2023). Even AI experts often struggle to grasp the mechanisms behind the resulting outcomes due to this complexity (Peters, 2023). In such scenarios, the reasoning behind the systems' decisions may not be readily apparent to users, nor those impacted by the systems (European Data Protection Supervisor, 2023).

2.3 Drug development

2.3.1 Original drug development processes

The average time it takes to bring a new drug to market is between two and twelve years, and the cost is prohibitively high (Tonkens, 2005). A new compound must cost roughly $1.8 billion to bring to market, and there are very few successful launches (Tonkens, 2005). The return on investment is this low because only one in 10.000 compounds reaches the market and of those, only one in three covers the cost of development (Tonkens, 2005).

In conclusion, the development of drugs is a difficult task (Tamimi & Ellis, 2009). Every product needs to be both safe and effective, and its effectiveness needs to be demonstrated across different ethnic, demographic and age groups (Tamimi & Ellis, 2009). The pharmaceutical sector is currently one of the most regulated in the world, and every drug must pass an international regulatory evaluation (Tamimi & Ellis, 2009). After this is completed, authorized items need to be appealing to international markets with diverse cultural backgrounds, healthcare systems and distribution networks (Tamimi & Ellis, 2009).

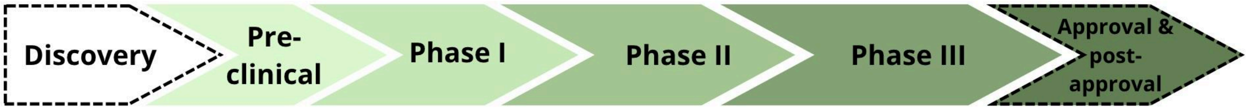


Figure 2.1: The traditional drug development process (Adapted from Tamimi & Ellis, 2009).

Discovery

Research and development in a lab setting is the first step in the discovery stage (see figure 2.1) (Biostock, 2023). The decision to invest in a certain area is influenced by the disease's prevalence and 'medical need' combined with other aspects such as R&D costs, technical feasibility, and commercial considerations such as rivalry among competitors and prospective market share (Tamimi & Ellis, 2009). The authors further point out that even if criterias are met, projects must compete for limited research and development funds, with the highest prioritized projects being selected for progression.

Preclinical research

Preclinical research, which can be done in vitro (in a test tube) or in vivo (in an animal), is required before any chemical is tested on humans with the reason behind preclinical research is mostly conducted to evaluate a compound's potential for serious damage (see figure 2.1) (Patheon, 2023). For instance, in the United States, the Food and Drug Administration receives an investigational new drug application (IND) asking for approval to start the strictly controlled process of clinical testing on human beings (Tonken, 2023). To sum up, the most crucial part of preclinical research is the stringent safety testing meant to make sure the candidate is safe before it is subjected to human clinical investigations (Biostock, 2023).

Biostock (2023), explains that once the preclinical tests are finished and the results support the researchers' ideas, the developers will request approval to move forward with clinical, or in-human, investigations. Either a Clinical Trial Application (CTA) in the EU or an Investigational New Drug (IND) application in the US is used to do this. After reviewing all the information, the relevant regulatory body determines whether to permit the move to the clinic (Biostock, 2023).

Clinical Drug Development

Phase 1

Following the regulatory approval and approval from ethics committees, phase I begins with the new medication being given to people for the first time (see figure 2.1) (Biostock, 2023; Tamimi & Ellis, 2009). While the preclinical research addresses fundamental concerns regarding the safety of a medication; it is not a replacement for research on the drug's potential interactions with the body (FDA, 2018).

With the exception of cytotoxic medications (such as cancer medicines), which are tested in patients without first needing to be examined in healthy volunteers, this phase typically involves healthy participants (Tamimi & Ellis, 2009). This phase aims to assess the tested drug's pharmacodynamic (i.e., how the drug affects the body, such as heart rate, blood pressure, electrocardiogram (ECG), etc.) and pharmacokinetic (i.e., how the body affects the drug, such as absorption, distribution, metabolism, and excretion) effects (Tamimi & Ellis, 2009). In this case, the candidate is often tested on 20 to 80 healthy volunteers to see if the results match the preclinical studies' findings regarding the candidate's behavior in the human body (Biostock, 2023). Of the drugs tested, approximately 70% move to the next phase (FDA, 2018).

According to Biostock, (2023). Phase 1 could be described as a safety phase, once more, the primary focus is on the substance's safety profile, or toxicity, but this time it is on humans. Phase I testing determines the safe dose, the drug's absorption profile, and the duration of the drug's activity in the body. It can take up to a year to complete a phase I study (Biostock, 2023).

Phase 2

Biostock, (2023) specifies that drug developers can request authorization to move on to phase II clinical development if phase I safety data are positive (see figure 2.1). During this stage, the candidate is typically assessed on 100 to 300 individuals who have been diagnosed with the illness the candidate is meant to treat (Biostock, 2023).

Phase IIa and IIb are the two typical divisions of phase II (Tamimi & Ellis, 2009). Phase IIa is the testing of the medicine in a small cohort (12–100 patients) as a "proof of concept," usually restricted to a single high/maximum tolerable dose level (Tamimi & Ellis, 2009). Phase IIb is the continuation of the proof of concept, wherein the target population is exposed to a range of dose levels (dose-ranging studies) in order to determine the minimally effective or non-effective dose and the best dose to proceed to the next stage based on safety and clinical efficacy (Tamimi & Ellis, 2009). Phases IIa and IIb are occasionally integrated into a single, sizable study (Tamimi & Ellis, 2009). As the drug's minimum and maximum dosages are established for usage in the following stage of development, efficacy and safety now come together. It can take up to two years for Phase II (Biostock, 2023). The phase 2 phase can take up to 2 years and only roughly 33% of the drugs selected for phase 2 passes to the next phase (Biostock, 2023) (FDA, 2018).

Phase 3

Phase 3 is the stage of a drug's evaluation before a request for market approval is made to pharmaceutical regulators (see figure 2.1) (Biostock, 2023). In order to receive approval, the clinical dosages, frequency, and timing of administration must be confirmed at this stage of drug development, which is the last one before registration (Tamimi & Ellis, 2009). Prior to starting an expensive phase III trial, the sponsor needs to be very confident about the drug's safety and effectiveness in the intended patient population and the dosage range that will be investigated (Tamimi & Ellis, 2009).

While phase III studies are primarily structured and powered to test the efficacy hypothesis, adverse events are also recorded to evaluate the drug's benefit-risk potential (Tamimi & Ellis, 2009). Regarding the phase III study, researchers record and disclose any adverse events that patients may have encountered (Biostock, 2023). This implies that in order to ensure that those side effects are appropriately evaluated, patients must be exposed to the medication for extended periods of time (Biostock, 2023). Any negative effects that are identified at this point are included in the final product's packaging leaflet (Biostock, 2023). The third phase takes on approximately one to four years and only 25-30% of the drugs that passed through each phase moves on to the next step of the drug development process (Biostock,2023; FDA, 2018). The entire success rate of phase III is roughly 70%, and the cost can reach USD 100 million depending on the size with a successful phase III being typically recognized by the financial markets, with an influence on the sponsor's stock price (Tamimi & Ellis, 2009).

Approval and Post Approval

Approval

If a drug developer has evidence from early tests, as well as preclinical and clinical studies, that a medicine is safe and effective for its intended use, it can submit an application to market the drug (FDA, 2018). That application is called an NDA (New Drug Application/ BLA (Biologics License Application) in the U.S. and MAA (Marketing Authorisation Application in the EU (Biostock, 2023). Thereafter a pharmaceutical authority such as EMA (European Medicines Agency), and 'Läkemedelsverket' thoroughly evaluates all submitted medication data and decides whether to approve it (FDA, 2018). The process of putting together an application combined with the process of an authority assessing the application could take all from 10 months to a few years (Biostock, 2023)

Biostock, (2023), clarifies that if regulatory authorities approve an application, the candidate, or medicine as it is now known, is ready for market release and currently the principal and possible buyers begin price discussions. The price bargaining procedure varies widely from country to country (Biostock, 2023). Biostock, (2023), explains that EU member states adhere to specific EU guidelines, but may also have country-specific norms. Further the authors bring up Sweden as an example where compensation policies are set on a national level. In contrast, in the United States, price talks between pharmaceutical corporations and private insurance companies take place without the involvement of the authorities (Biostock, 2023). The method has resulted in much higher medicine prices in the US compared to Europe and other industrialized countries (Biostock, 2023).

Post Approval

This phase, often known as phase 4 studies, includes regulatory agencies requiring phase IV research after a medicine has been approved for commercial use (Biostock, 2023). This is accomplished by gathering data from clinical practice - that is, from actual care units that treat patients (Biostock, 2023). Even while clinical trials provide essential information about a drug's efficacy and safety, all information on a drug's safety cannot be obtained at the time of licensure (FDA, 2018). Despite the rigorous stages involved in medication research, there are limitations (FDA, 2018). As a result, the full picture of a product's safety emerges during the months and even years that it spends in the marketplace (FDA, 2018).

Pharmacovigilance is defined as the science and activities related to discovering, evaluating, understanding, and preventing side effects of drugs, as well as all other drug-related problems (World Health Organization, 2024). In Europe, pharmacovigilance is used to describe the ongoing review of the safety of the drug in the post-marketing era; it is a requirement that all pharmaceutical companies having a post-marketed product must fulfill (Tamimi & Ellis, 2009). The goal is to promote pharmacovigilance which (Biostock, 2023). Phase IV investigations determine whether the medicine interacts with other substances, and additional safety testing is performed (Biostock, 2023). The authors further explain that this is especially essential for medications used to treat complex medical diseases or pregnant women who are unlikely to have participated in phase I-III studies.

Conclusively BioStock, (2023), disclose that phase IV studies may be useful for medications that treat rare illnesses, which often have a small number of patients in phases I-III. The results of the prior clinical studies thus have lesser statistical certainty, which is why the regulators are asking for further validation of the safety and efficacy of the medication (Biostock, 2023).

2.3.2 Drug Development with Artificial Intelligence

The application of AI in the medicine sector has received substantial interest in recent years as a potential means of changing the pharmaceutical industry (Paul et al., 2021). Drug discovery, or the process of discovering and developing novel treatments, is a complicated and time-consuming undertaking that has traditionally relied on labor-intensive procedures such as trial-and-error testing and high-throughput screening (Blanco-González et al., 2023). It is estimated that every new drug introduced to the market costs billions of dollars and takes more than ten years to develop (Chen et al., 2021). As a result, measures that help facilitate and accelerate the drug development process are highly desirable (Chen et al., 2021).

Recent years have witnessed significant strides in drug development and discovery techniques (Tripathi et al., 2021). Various companies have leveraged advancements in AI and machine learning technologies, affording them competitive edges (Ruedig & Kler, 2022). Moreover, firms specializing in AI are capitalizing on their expertise by offering consultancy services and technology to other companies (Ruedig & Kler, 2022).

Due to the current competitive nature and growth of new technologies in the area, there are too many adaptations of AI in drug development to cover in this section. However, examples of how AI is currently used to facilitate and improve drug development processes are presented below:

Drug Discovery

According to Davids et al. (2022) the initial stages of drug development, the selection of compounds, is one of the most challenging ones. In order to operate effectively in this drug discovery phase, machine learning algorithms are necessary. They present the information that there are many ways for using AI in drug discovery and development, although they highlight the limitation due to the black box. Many of the potential uses discussed by Davids et al. (2022) are also in need of more research before they are put into practical use.

High Throughput Screening (HTS) techniques and computational chemistry have advanced throughout time, accelerating the timely screening of millions of compounds against the precise identified diseases (see figure 2.2), (Tripathi et al., 2021). These methods make use of enormous amounts of biological data, bringing drug discovery into the big data era. Big data analytics is now used outside of information technology to impact several industries, including drug discovery (Tripathi et al., 2021). It is extremely important to mine this complicated and heterogeneous data across numerous resources (Tripathi et al., 2021). As a result, the research communities now face both obstacles and opportunities as they develop new computational tools and algorithms for managing and curating large amounts of data (Roy et al., 2010).

Furthermore, the development of AI and machine learning (ML) algorithms, along with improvements in high computing facilities, is crucial and have been applied to computer-aided drug design technology because it allows lead-like molecules to be mined and screened more quickly and efficiently against the intended target (Sliwoski et al., 2014). As illustrated in figure 2.1 AI also helps to identify possible medicinal molecules faster and more efficiently than traditional approaches (Tripathi et al., 2021).

Drug Testing Phases

AI systems possess the capability to imitate and forecast how different substances interact with biological targets, thereby reducing the duration and expenses associated with pharmacological trials (Tripathi et al., 2021). This proficiency becomes particularly crucial in swiftly addressing emerging global health challenges, such as creating vaccines and treatments for novel diseases (Tripathi et al., 2021).

According to Paul et al. (2021) the conventional success rate of drugs progressing into the clinical trial stage stands at a mere 10%, indicating that companies invest significant time and resources in lengthy testing phases that often culminate in failure. They point out that the underlying reasons for the high rate of unsuccessful clinical trials typically stem from issues like inadequate patient selection, substandard infrastructure, or insufficient technical prerequisites. Paul et al (2021) continues to explain that these setbacks can be mitigated through the utilization of predictive machine learning to evaluate the potential efficacy of drugs before their initiation into clinical trials, thereby saving both time and resources by eliminating drugs predicted to be ineffective (see figure 2.2). Additionally, AI assistance in patient selection and monitoring can bolster the success rate of clinical trials (see figure 2.2) (Paul et al., 2021). Up to 30% of clinical trial failures result from patient dropouts, a predicament that AI monitoring and management can alleviate (Paul et al., 2021).

According to industry experts, the significant failure rate in drug development is cited as a key factor driving high prices for medicines and vaccines (Paul et al., 2021). By reducing the frequency of failures, it is believed that the drug market could become more affordable (Paul et al., 2021).

Post Approval

As stated by the World Health Organization, (2024), the pharmaceutical corporations are increasingly turning to AI for managing the distribution and supply chain of medications (see figure 2.2), including monitoring the integrity of vaccine transportation. AI proves invaluable in predicting demand, identifying irregularities in the supply chain, and foreseeing potential shortages and stock deficits (World Health Organization, 2024). Additionally, the authors illustrate that AI plays a pivotal role in devising marketing strategies, determining pricing strategies, and pinpointing optimal sales outlets for newly developed drugs. By analyzing various data sets such as demographics, medical records, prescriptions, and feedback from healthcare providers, AI aids in customizing marketing tactics, utilizing channels like email campaigns, advertisements, and social media to effectively engage with healthcare professionals (World Health Organization, 2024). Generative AI, like large language models, finds utility in patient education and fostering medication adherence (see figure 2.2), potentially integrated into platforms for patient support or digital therapeutic applications (World Health Organization, 2024).

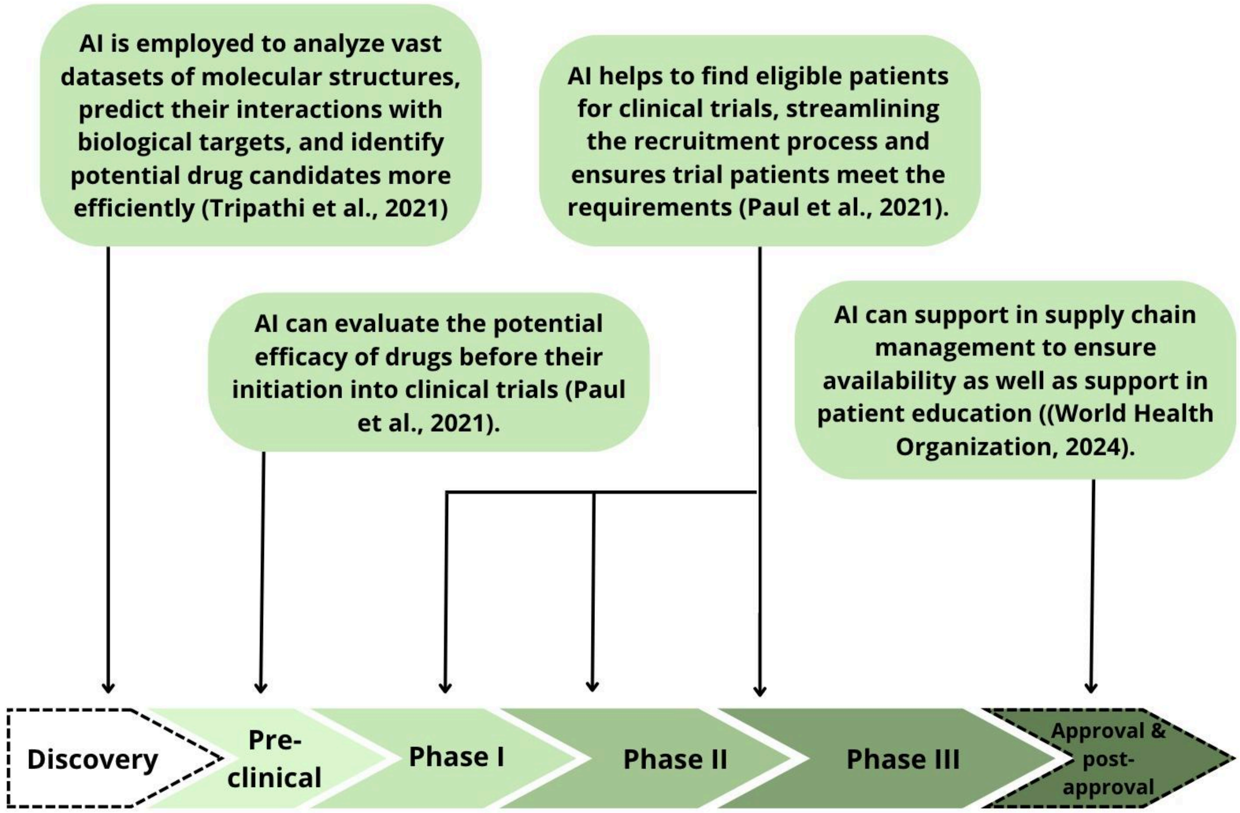

AI is employed to analyze vast datasets of molecular structures, predict their interactions with biological targets, and identify potential drug candidates more efficiently (Tripathi et al., 2021)
AI helps to find eligible patients for clinical trials, streamlining the recruitment process and ensures trial patients meet the requirements (Paul et al., 2021).
AI can evaluate the potential efficacy of drugs before their initiation into clinical trials (Paul et al., 2021).
AI can support in supply chain management to ensure availability as well as support in patient education ((World Health Organization, 2024).
Discovery
Pre-clinical
Phase I
Phase II
Phase III
Approval & post-approval

Figure 2.2: Visual representation of examples where AI can benefit the drug development lifecycle (Adapted from Tamimi & Ellis, 2009; modifications and comments added by the authors).

2.4 Research Frameworks

2.4.1 Explainable Artificial Intelligence (XAI)

In recent years, the advancements of AI technologies has sparked both excitement and apprehension regarding their widespread adoption across various domains (European Commission, 2020). The opacity of AI systems, particularly in their decision-making processes, is one major source of concern (European Commission, 2020). To address this challenge, 'Explainable Artificial Intelligence (XAI)' has become a topic of interest for research and development because XAI systems objective is to offer an explanation for humans that is easier to interpret and understand (European Commission, 2020). According to Gunning et al. (2019) the fundamentals of XAI systems is that the system should be capable of describing its understandings and capabilities, as well as what it has done, is doing, and will do in the future. It may also reveal the important information that it is acting upon. The main focus of such a system is thus on the reasoning behind the prediction or decision of an AI-system (Gunning et al., 2019; Phillips et al., 2021).

As Hassuja et al. (2024) states, the phrase ‘black box’ in XAI is frequently used to contrast with "white box" or "transparent" models, where the internal logic and workings of the predictions are readily available and understandable. In general, it aids users in fully comprehending and having faith in the choices these systems make (Hassija et al., 2024). Highly effective prediction models, like DNNs, which are a type of machine learning algorithms that, like artificial neural networks, try to emulate the brain's information processing, typically have some inherent transparency flaws that must be fixed to support their application in a variety of settings (Hassija et al., 2024). Systems can be opaque to the point where a human being does not know elements that are epistemically relevant; therefore, the degree of opacity may differ based on the agent (Humphreys, 2009). Generally speaking, though, machine learning algorithms can be opaque in two senses: (1) the way that machine learning produces outputs from certain inputs may be opaque or difficult to understand, and (2) programmers or observers may not be aware of the inputs themselves (Burrell, 2016). This opacity may result from machine learning features, proprietary concerns, or technical illiteracy (Burrell, 2016).

However, according to Miller (2017), rather than successfully meeting the needs of the intended users, many efforts to improve explainability frequently result in explanations that are primarily suited to the AI researchers themselves. This puts the onus of developing a compelling explanation for intricate decision models on AI specialists who possess a thorough comprehension of these models (Miller, 2017). Explaining the system's understandings and abilities, as well as its previous acts, current procedures, and future moves, and disclosing the pertinent data that informs those actions are all ideal components of XAI (Gunning, 2019).

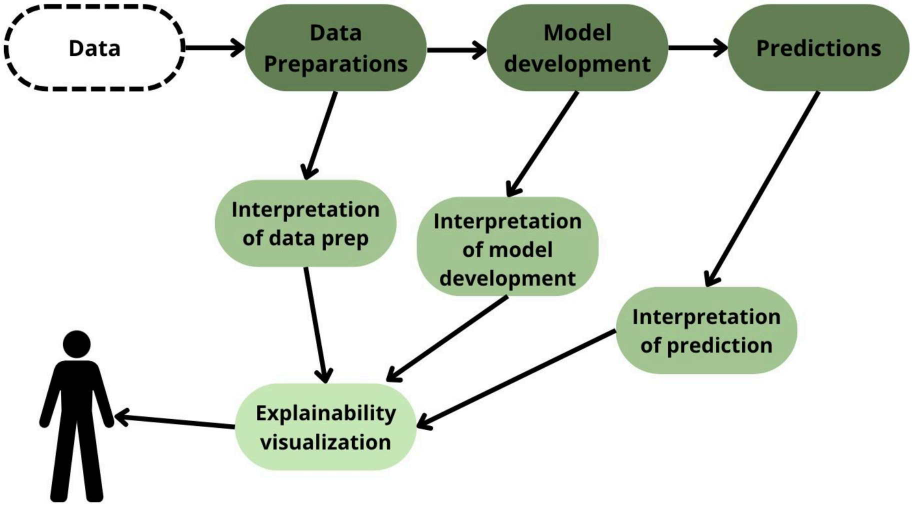

Data
Data Preparations
Model development
Predictions
Interpretation of data prep
Interpretation of model development
Interpretation of prediction
Explainability visualization

Figure 2.3: Visual representation of the concept of XAI (Adapted from McDermid et al., 2021).

The landscape of Explainable Artificial Intelligence (XAI) is shown in Figure 2.3. The cycles of data processing, outcome prediction, model training, and interpretive feedback are depicted in the picture (McDermid, 2021). These cycles are all connected by explainability (McDermid, 2021). This framework's components are all intended to foster comprehension and clarity, making it possible for all parties involved to utilize AI systems' functions (McDermid, 2021). Explainability is based on the idea that complex AI processes can be made simple enough to follow and comprehend, allowing inputs to be converted into valuable outputs (McDermid, 2021).

Figure 2.3 serves as an educational and operational guide in the application of XAI, detailing how various components such as data collection, algorithm development, and model evaluation interlink to foster an environment of transparency (McDermid, 2021). By visually breaking down these interactions, Figure 2.3 highlights the importance of incorporating explainability throughout the AI model lifecycle, and how this is done with the XAI concept (McDermid, 2021). McDermid further discusses that this strategy guarantees that the reasoning behind decisions made by AI is transparent and defensible, while also improving confidence and accountability in AI systems. In order to ensure that AI systems are both efficient and ethically sound, such clarity is essential in encouraging the adoption and responsible usage of these technologies across a range of fields (McDermid, 2021).

2.4.2 The Assessment List for Trustworthy Artificial Intelligence (ALTAI) by EU

According to the European Commission (2020), the main purpose of the ALTAI is to help organizations handle their use of AI, understand what trustworthy AI is and what risks there are. While still maximizing the potentials and benefits of AI, ALTAI is founded in people's fundamental human rights and the protection of these (European Commission, 2020). The European Commission, (2020), points out that generally, within the EU, the list is intended to be used for self-evaluation while utilizing AI, not just for medical or healthcare purposes. The list includes seven criteria in all: (1) Human agency and oversight; (2) Robustness and safety of the technology; (3) Privacy and Data Governance; (4) Transparency; (5) Diversity, Non-Discrimination and Fairness; (6) Social and Environmental Well-Being; and (7) Accountability (European Commission, 2020).

All of the above stated requirements are relevant för AI use, however, the guide is created so that an organization takes into account the requirements relevant for the specific AI system they are utilizing (European Commission, 2020). For the sake of this research, we will be considering the requirements for ALTAI that regards the topic of explainability issues of Black Box AI models.

We acknowledge that all requirements relate to this topic in some way, however, we will be focusing on the fourth requirement regarding transparency. This study is researching specifically the area of explainability and Black Box models. It is in the section Transparency that explainability issues are addressed.

Requirement #4 Transparency

In the list, the transparency factor in the list has three subsections which this requirement is meant to cover, the three are traceability, explainability and open communication about the limitations of the AI system. We will be including explainability in this section as it is the subsection directly relevant to explainability of black box models in the drug development industry as well as the XAI theory.

Explainability

The European Commission (2020), defines explainability in this subsection of the requirement as the capability of explaining the technical as well as the reasoning leading up to the decisions or predictions made by the AI system. However, they acknowledge that total explainability is not always possible in the case of black boxes. These cases are to be paid special attention to and other explainability measures can possibly be used, the ones mentioned are traceability, auditability, and transparent communication on the abilities of the AI (European Commission, 2020). To what extent the explainability is needed is determined by how critical a failed output would be to a human life (European Commission, 2020). However, the list does not give any further indication of how such systems are handled in different situations.

2.4.3 Artificial Intelligence Act (AI Act) by EU

The European Union's AI Act is a regulatory framework aimed at ensuring that AI systems are transparent, lawful, and fair, focusing particularly on high-risk and safety-critical applications (European Parliament and Council of the European Union, 2024). It establishes uniform rules for the development, marketing, use, and monitoring of AI systems within the EU, promoting the uptake of human-centric and trustworthy AI while ensuring a high level of protection of health, safety, and fundamental rights as outlined in the Charter of Fundamental Rights of the European Union (European Parliament and Council of the European Union, 2024).

The AI Act mandates that high-risk AI systems must be designed in a manner that operations can be understood and traced by humans, with clear documentation of algorithms' functioning and decision-making rationales to ensure accountability and oversight (European Parliament and Council of the European Union, 2024, 2024).

The European Parliament and Council of the European Union (2024) outlines specific requirements for high-risk AI systems to comply with fundamental rights, non-discrimination, and consumer protection laws. It emphasizes the need for AI systems to be secure, error-free, and robust against attempts to manipulate them. Furthermore, all AI systems are required to be transparent about their capabilities and limitations, and humans must always be aware when they are interacting with an AI system (European Parliament and Council of the European Union, 2024).

For high-risk AI applications, the document specifies that these systems must undergo a conformity assessment to verify compliance with the regulation's requirements before being put on the market or put into service. It also addresses the need for adequate data governance and management practices to support the functioning of AI systems (European Parliament and Council of the European Union, 2024).

AI Act Risk Levels

The AI Act defines four different risk levels when it comes to AI systems, each requiring different assessment and examination (European Commission, 2024). According to the European Parliament (2023), medical devices are included in the high-risk category of the AI Act risk assessment scale. According to The World Health Organization (WHO) (n.d.), the term medical device includes, among other things, software and machines intended to be used for medical purposes.

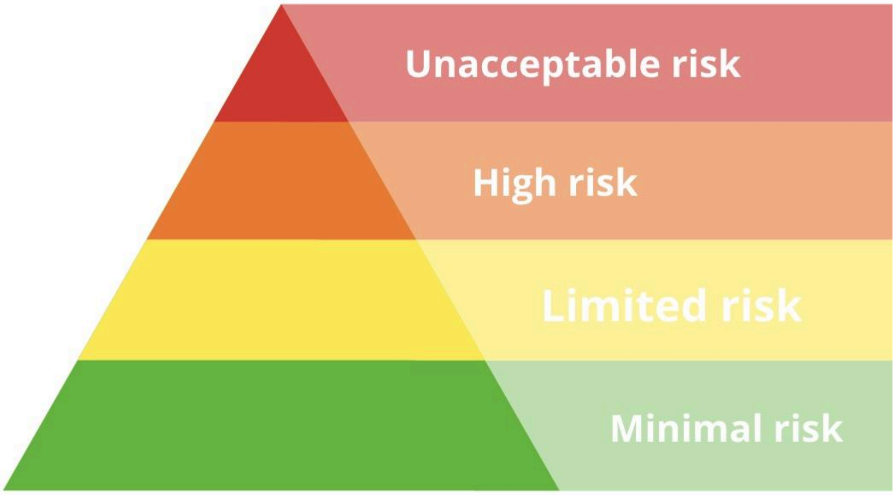


Figure 4: AI Risk Level Framework (Adapted from European Commission, 2024).

Minimal Risk

Most AI systems fall under this category where the AI Act does not impose any new legal requirements (see figure 2.4) (European Commission, 2024).

Limited Risk

These AI systems require specific transparency obligations. An example is chatbots, where users should be aware that they are interacting with an AI (see figure 2.4) (European Commission, 2024).

High Risk

According to the European Parliament (2023), among others, all AI systems integrated into products governed by EU product safety legislation are classified as high-risk, including those used in medical devices (see figure 2.4) (European Parliament, 2023). Such systems are subject to stringent compliance standards, which encompass the use of high-quality datasets, comprehensive documentation, traceability, human oversight, and robustness (European Commission, 2024).

Unacceptable Risk

Certain uses of AI are seen as creating risks that are unacceptable and are banned (see figure 2.4) (European Commission, 2024). This includes AI that manipulates human behavior to circumvent users' free will, and systems that allow 'social scoring' by governments (European Commission, 2024).

2.5 Previous Research

2.5.1 EU and the XAI Theory (Projects XAI and NL4XAI)

Between 2019 and 2025, the European Commission, (2024) has been actively investing in projects aimed at advancing Explainable Artificial Intelligence (XAI) and enhancing the transparency of black box systems. In 2020, the Commission initiated two key projects, namely XAI and NL4XAI, both centered around the foundational principles of XAI (Cassauwers, 2020). The NL4XAI project, slated for completion in September 2024, and the XAI project, scheduled to conclude in September 2025, represent significant strides in addressing the challenges posed by opaque AI systems (Cassauwers, 2020; European Commission, 2024). These initiatives underscore the EU's commitment to leveraging XAI as a tool to mitigate concerns surrounding the use of AI.

The scientific coordinator of NL4XAI, (2020), outlined two potential solutions in 2020. The first involves the development of transparent "white box" models, while the second entails creating algorithms capable of translating the inner workings of black box systems into comprehensible natural language (Cassauwers, 2020). However, the coordinator also emphasized a critical issue associated with the second approach: determining what constitutes an adequate explanation and establishing criteria for acceptable explanations (Cassauwers, 2020). This highlights the challenges inherent in the pursuit of XAI and underscores the importance of comprehensive research and interdisciplinary collaboration in this domain.

2.5.2 European Data Protection Supervisor (EDPS); TechDispatch

TechDispatch reports are frequently released by the European Data Protection Supervisor, (2023), (EDPS) with the goal of elucidating the latest technological advances. The TechDispatch reports are a component of the broader technology monitoring initiatives of the EDPS (European Data Protection Supervisor, 2023). Every TechDispatch offers accurate summaries of newly developed technologies, offers connections to more suggested reading, and offers a preliminary evaluation of any potential effects on personal data protection and privacy (European Data Protection Supervisor, 2023).

The EDPS, (2023), publishes TechDispatch reports every year, which are forums for debating the most recent developments in technology. These studies are important in today's technological world because they address a wide range of subjects with this year's TechDispatch discussing the usage of AI across industries and the importance of AI Explainability. The TechDispatches report does not only remain one updated on the latest advancements in AI applications and the need for explainable AI but also gives valuable insights into the ethical considerations concerning Black Box models and regulatory frameworks concerning explainable AI in the EU (European Data Protection Supervisor, 2023).

Transparency

The EDPS (2023), states that a transparent AI system fosters accountability by enabling stakeholders to validate and audit its decision-making processes. It facilitates the detection of biases or unfairness and ensures that the system operates in accordance with ethical standards and legal requirements (European Data Protection Supervisor, 2023).

Lepri et al (2018), further explain that transparency entails the capacity for a particular model to be comprehensible and in its strictest sense, a model is transparent if an individual can grasp its entirety simultaneously. Transparency can be evaluated at various levels: the complete model level, the level of individual components (such as parameters), and the level of a specific training algorithm (Lepri et al., 2018).

A less stringent interpretation of transparency suggests that each aspect of the model (e.g., inputs, parameters, and computations) allows for an intuitive explanation (Lepri et al., 2018).

Interpretability

Interpretable AI models enable humans to predict the outcome a model will generate given an input, and to discern when the model has made a mistake (European Data Protection Supervisor, 2023)

The term "interpretability" describes how easily a certain "black box" model or conclusion may be understood by humans (Miller et al., 2017). Poorly interpretable models are opaque because they often leave individuals without a clear understanding of how or why a specific classification has been reached based on the inputs provided (Burrell, 2016).

Explainability

Explainability holds significant importance in critical applications involving human lives or sensitive information, as it aids users, regulators, and stakeholders in understanding the reasoning behind AI-generated outcomes (European Data Protection Supervisor, 2023).

Thampi, (2002), describes that explainability in AI focuses on delivering transparent and coherent explanations for individual model predictions or decisions. Its goal is to address queries like "Why did the AI system produce this specific prediction?" by providing human-comprehensible justifications or rationales for particular outcomes. The authors point out that while built on interpretability, explainability also draws insights from diverse disciplines such as human-computer interaction, law, and ethics (Thampi, 2002).

2.5.3 XAI in the AI Act

According to Panigutti et al.'s (2023) interpretation, the AI Act does not mandate the complete replacement of 'black-box' AI models with entirely transparent ones but encourages the use of XAI techniques where possible to enhance transparency and human oversight. Their study describes how human oversight should be managed, referring to the AI Acts requirements that a natural person (human) should have necessary understanding, authority and training of the system (Panigutti et al. 2023). Furthermore, they go on to describe that a human needs to be able to control the system and understand its risks and potential bias. These requirements do not represent all requirements for high-risk AI systems by the AI Act, but they outline a human-centric approach interpreted by Panigutti et al. (2023). This suggests a balanced method, promoting safety and accountability without stifling innovation. Panigutti et al. (2023) highlight that this approach is needed especially in the case of high-risk AI models, such as medical devices (European Parliament, 2023).

2.5.4 Reflections by the European Medicine Agency on EU requirements for Artificial Intelligence in Medicine

The European Medicines Agency (EMA), is a decentralized entity of the European Union, tasked with the scientific assessment, oversight, and safety regulation of pharmaceuticals intended for the EU market (European Medicine Agency, 2023). Established in 1995, EMA ensures the protection of public and animal health within EU and European Economic Area nations by guaranteeing the safety, efficacy, and quality of all medicinal products available (European Medicine Agency, 2023).

EMA (2023) has presented a draft on AI use in the context of the medicine lifecycle. It outlines potential risks associated with AI/ML and the necessity for rigorous standards of explainability, particularly reflecting concerns like those addressed by EU regulations (European Medicine Agency, 2023).

EMA (2023) highlights one example in the text which concerns the use of AI/ML in generating medical research outputs. The draft stresses that while AI tools can efficiently process data and generate insights, they also pose risks such as the potential of biases and misleading medical advice. This is particularly risky in medicine, where inaccurate information can lead to inappropriate treatment decisions (European Medicine Agency, 2023). To mitigate these risks, the document advocates for clear disclosure of AI involvement in research processes, which aligns with EU regulations on AI transparency and explainability (European Medicine Agency, 2023). These regulations require that AI systems used in healthcare be designed in such a way that their decisions can be understood and traced by human operators, ensuring that outputs are verifiable and scientifically sound (European Medicine Agency, 2023).

Additionally, the European Medicine Agency, (2023) discusses the ethical implications of AI-generated content in medicine, emphasizing the importance of maintaining research integrity. They mention the need for human researchers to take full responsibility for the content produced with AI assistance, ensuring that all medical recommendations or conclusions derived from AI analyses are critically reviewed and validated by human experts. This approach is crucial in medical contexts where the stakes are high, and the impact of misinformation can be severe (European Medicine Agency, 2023). By insisting on human accountability, the document aligns with the EU's emphasis on maintaining human oversight over AI, ensuring that AI tools are used to enhance, rather than replace, human expertise in critical decision-making processes (European Medicine Agency, 2023).

The European Medicine Agency, (2023), advocates for a balanced approach where AI aids medical research without compromising the quality and integrity of the scientific process, ensuring that all AI-generated medical advice is transparent, understandable, and, most importantly, verifiable. They advise a human-centric approach in all use of AI and ML in the medicine sector (European Medicine Agency, 2023).

2.5.5 Guidelines for Artificial Intelligence in Swedish Healthcare

The Swedish Medical Products Agency, Läkemedelsverket, has produced a guide that can be used as a reference when integrating and utilizing AI in the healthcare system (Läkemedelsverket, 2024). This is the first Swedish guide on AI applications in healthcare and medicine, and it was released in November 2023. They expect that for the foreseeable future, striking a balance between the requirement for basic clinical validation and technological innovation will be a significant problem. This report has been designed in accordance with EU ALTAI requirements for ethical use of AI (Läkemedelsverket, 2024).

The Swedish legislation includes equality in the context of healthcare, making any use of biased or unequal AI models in the medical field illegal (Läkemedelsverket, 2024). Because of this, it is crucial that the datasets used to train machine learning algorithms be entirely unbiased. Any application of AI and ML in the healthcare industry, according to Läkemedelverket, should assess the possibility of discrimination and handle any potential bias in the system using manual processes or alternative systems (Läkemedelsverket, 2024).

Due to their reduced bias and ability to be monitored, Läkemedelsverket, (2024) strongly advises using transparent and explicable AI models. They do concede, though, that some of the best machine learning systems—like deep neural networks—operate as "Black Box" systems. In their research, Läkemedelsverket, (2024), evaluates these systems and explains that, if it can be demonstrated that no other transparent AI model can match these models' performance and cutting-edge outcomes, then the usage of these Black Box systems in the healthcare industry may be justified (Läkemedelsverket, 2024).

2.5.6 Developers on Ethical Artificial Intelligence in Medicine Applications

Designing AI applications for medicine is a critical development process considering the ethical aspects that need to be considered (Weidener & Fischer, 2024). Weidener and Fischer's study, (2024), explores the use of AI in medicine, particularly in medication development, medical research, and other areas. Investigating the opinions of AI experts on the moral use of AI and the connection between morality and the further advancement of AI in medical applications is the main goal of their study (Weidener & Fischer, 2024).

The study showed three main themes in the experts' answers, including seven subcategories within the three themes (Weidener & Fischer, 2024):

Essential Foundation

Weidener and Fischer, (2024), found that the the essential foundation included the subcategories "Awareness", "Consequences" and "Data protection".This category gave perspective of a consensus that ethics in AI is important and that developers do need to acknowledge it due to risks and consequences (Weidener & Fischer, 2024).

Awareness refers to the situations where AI is used in ways that were not anticipated by the developer, and that the developer needs to be aware of this (Weidener & Fischer, 2024). Consequences covers the risk of bias and discrimination in an AI model, and it is highlighted that this is something that everyone in every team, or anyone at all working with AI, needs to consider (Weidener & Fischer, 2024). Data protection refers to safe handling of human-data when developing medical applications (Weidener & Fischer, 2024).

Results in the Foreground

Under the "Foreground," the subcategories "Performance" and "Economic Efficiency" represent aspects indicating a results-oriented approach in the development process of AI models (Weidener & Fischer, 2024). Performance emerges as a recurring pattern among multiple experts, who prioritize dealing with the technical efficacy of the system over ethical considerations during development (Weidener & Fischer, 2024). This underscores the emphasis on performance metrics rather than ethical dimensions in the development process (Weidener & Fischer, 2024).

Closely intertwined with the Performance subcategory, Economic Efficiency reflects the financial priorities of AI development companies, which are often reliant on funding (Weidener & Fischer, 2024). This dependency may sometimes lead to compromises in ethical standards in favor of financial gains (Weidener & Fischer, 2024).

Obstacle to progress

This category includes subcategories 'Demands' and 'Blockade', which represent the recurring statements about ethics posing as an obstacle for technological progress (Weidener & Fischer, 2024). Demands refer to ethics as being a barrier for technological progress and AI potential (Weidener & Fischer, 2024). The uncertainty is highlighted as experts find it hard to know where to draw the line to keep the AI ethical and still maximizing the potential of the system (Weidener & Fischer, 2024). Blockade, closely aligned with "Demands," pertains more to future implications (Weidener & Fischer, 2024). Here, experts view ethics not only as a barrier to current AI utilization but also as a hindrance to the long-term evolution of AI (Weidner & Fischer, 2024).

Weidener & Fischer's study (2024) concluded that though there is a general understanding that ethics are important, to what extent the respondent prioritizes the ethical aspects seem to vary.

2.5.7 Healthcare Leaders on Artificial Intelligence in Healthcare

A study published in BMC Health Services Research (2022) delves into the perspectives of healthcare leaders in Sweden regarding the integration of AI in their field. The research aims to uncover the obstacles these leaders perceive in the advancement of AI within healthcare. Participants in the study encompass various roles and professions within the healthcare sector, although not specifically drug developers (Petersson et al., 2022). The authors point out that while the study predominantly reflects a managerial viewpoint on AI implementation in healthcare, healthcare leaders do raise concerns about explainability, and ethics related to AI.

Leaders emphasize the critical importance of addressing ethical challenges, particularly when AI applications may impact individuals' health within the healthcare sector (Petersson et al., 2022). Issues surrounding the reliability and accuracy of AI systems are raised, with specific mention of the "black box" problem. One leader underscores the necessity for clear explanations of AI systems to the users (Petersson et al., 2022). In essence, the study underscores the significance of ethical concerns related to opaque AI systems as perceived by healthcare leaders (Petersson et al., 2022).

While ethics took center stage in the study, leaders also expressed concerns about the outdated nature of regulations and legislation governing AI usage, given its dynamic evolution (Petersson et al., 2022). This dual observation suggests a call for improved regulatory frameworks or highlights how current regulations may stifle AI advancement.

In an article by Furlong (2023), she presents the opinions of Andrew Hopkins, a prominent figure in AI-driven drug development, who echoes the previous sentiment about regulating AI. He warns against excessive regulation in the nascent stages of AI integration into drug development, emphasizing the importance of first understanding how AI will be utilized in the industry. Hopkins also points out the existing layers of regulation in drug development concerning safety and privacy. Another voice at the forefront of AI in the drug development sector, Jim Weatherall, urges regulators to work more progressively when it comes to regulating AI use (Furlong, 2023).

Iniesta (2024), a statistics expert in precision medicine, emphasizes the significance of prioritizing explainable AI when integrating AI into patient care. Additionally, she underscores the importance of respecting individuals' rights to consent to AI-driven decisions regarding their medical conditions. This means that if a patient opposes the use of AI in addressing their medical issues, their decision should be honored and respected (Iniesta, 2024).

2.5.8 Medical Technology Industry Perspective on AI Act

MedTech Europe's (2024) detailed response to the final AI Act provides valuable insights into the medical technology industry's perspective on regulatory challenges, especially regarding the integration of AI within European healthcare systems.

Ethical Concerns and Regulatory Alignment

MedTech Europe (2024) highlights concerns regarding the alignment of the AI Act with existing regulations like the Medical Devices Regulation (MDR) and the In Vitro Diagnostic Medical Devices Regulation (IVDR). These regulations ensure that medical technologies, including AI, follow strict guidelines for performance and safety. The organization points out potential conflicts between overlapping regulatory frameworks, which could lead to

uncertainties and hinder the timely delivery of innovative medical solutions. This concern mirrors the ethical issues in drug development, where clarity and transparency in AI applications are essential to maintaining trust and accountability (MedTech Europe, 2024).

Recommendations for Regulatory Clarity

To address these challenges, MedTech Europe (2024) proposes the development of clear European Commission guidelines and active stakeholder involvement, ensuring that AI regulations are well aligned with the practical realities of medical technology and drug development. The organization advocates for unifying AI-specific regulatory requirements with existing regulations to eliminate differences and redundancies that could compromise patient safety and product safety (Medtech Europe, 2024).

2.5.9 Epistemological and Ethical Issues in the Trust of Black Box Algorithms in Medicine

Durán and Jongsma (2021) examine the complex problems related to the reliability of black box algorithms in the medical domain, including both ethical and epistemological considerations. They start addressing the opacity characteristics of these algorithms, which often leads to a lack of transparency about how decisions are derived. This opacity raises concerns about the ability to verify and validate the processes behind algorithmic decisions, which is vital for trust in medical AI (Durán & Jongsma, 2021).

Durán and Jongsma (2021) introduce an approach termed “computational reliabilism” which asserts that trust in Black box algorithms should be based on their reliability rather than their transparency. According to this view, if an algorithm consistently produces reliable outcomes, it may be deemed trustworthy, even if its internal workings are not fully understood or transparent.

This shifts the focus from a traditional emphasis on transparency to an emphasis on outcome reliability as a basis for trust (Duraá & Jongsma, 2021).

Further, Durán and Jongsma (2021) explore how this shift impacts ethical considerations such as accountability and patient autonomy. They point out that while algorithms may be reliable, the ethical implications of their use must still be addressed, particularly in terms of how decisions are made and the implications of these decisions on patient care. They argue that ethical issues remain prominent, especially when decisions made by opaque algorithms affect patient outcomes (Duraá & Jongsma, 2021).

In discussing the role of physicians in using these algorithms, the authors note that medical professionals must navigate the balance between relying on the AI's recommendations and maintaining their professional responsibility for patient care (Durán & Jongsma, 2021). This involves understanding when and how to trust the algorithm's outputs and integrating them into the broader clinical context (Durán & Jongsma, 2021).

3 Methodology

Our methodology chapter covers our choice of research method, how collection of data has been conducted and analyzed. It is also presented what ethical considerations have been made and how we validate our research method.

3.1 Choice of Method

The foundation of our study was laid through an extensive review of existing literature in the relevant field. We examined various methodologies employed by other scholars and the outcomes they had in similar contexts. This analysis aimed to equip us with a broad comprehension of the subject matter before conducting interviews. Such understanding not only informed our approach but also played an important role in crafting good interview questions for the following qualitative interview study.

The literature review drew from journals, articles, and books discussing Artificial Intelligence (AI) within drug development and ethics. Prior research contributing to our study was limited to data up until 2021. We deemed any data, besides data regarding our theory section, collected before 2021 to be outdated due to the rapidly evolving nature of the field.

Through interviews, we sought to capture rich, detailed insights from participants, allowing us to gain a profound understanding of their perspectives, experiences, and interpretations. This qualitative approach was deemed ideal for exploring the intricacies of the topic, offering valuable qualitative data that could not be captured through quantitative methods alone (Oates, 2006).

3.2 Literature Search

Keeping our findings relevant amidst the fast-paced evolution of AI and ML technologies presents a challenge. The constant influx of new developments and advancements can quickly make aspects of our research obsolete, requiring ongoing updates and revisions to stay current. Despite restricting our literature review of prior research to works authored no earlier than 2021, there's still a risk of encountering outdated data. There were instances when older sources were used in our research, although not in research of modern drug development and AI advancements.

Previous research and contents of the literature review in this study have been collected from journals, articles and existing studies on related and similar topics, as well as information which has helped us in defining terminology. To ensure the credibility of our research, we primarily utilized LUBSearch, Lund University's online library, as our main search engine, supplemented by Google Scholar. The following are the primary search terms used to gather relevant literature:

- Artificial Intelligence
- Black Box
- Machine Learning
- Opaque Systems
- Explainable Artificial Intelligence (XAI)
- Ethical AI
- Drug Development
- Drug Development processes
- EU AI
- European Commission AI
- European Parliament AI
- Pharmacovigilance AI
- Pharmaceutics AI

Search terms were used in different combinations and orders with linking words as "in" and "on" as well as "attitudes", "perspectives" and "views".

The collection of literature gave us a foundational understanding of the area and was thus what shaped our research question and our research theory. What we gathered from our initial readings of literature was that requirements for Black Box AI are rather diffuse and there seems to be room for individual interpretations of the guidelines.

3.3 Empirical Data Collection

3.3.1 Setting

The participants in this interview are all Sweden based, which is a product of our possibility to connect with relevant candidates for interviews. However, Sweden is one of the top 5 countries within the field of drug discovery and development as of 2023 (SwedenBIO, 2023).

Furthermore, the number and accessibility of participants for interviews may have constrained the diversity and depth of perspectives captured in our study. This limitation in data availability may impact the comprehensiveness and representativeness of our findings, particularly in reflecting the full spectrum of perspectives of the drug development industry. Additionally, there is a chance that respondents with AI knowledge who were confident enough in the subject to take part in our interview and were comfortable enough to discuss AI in drug development procedures would perceive the technology more positively or negatively than someone who is less knowledgeable about it.

3.3.2 Respondents

Participants in this interview study have been selected based on their work and knowledge within the AI in healthcare and drug development. Through our research of the topic, we have come across individuals who either work in an organization which is active within the field of research or are researchers studying the topic.

We found that an effective way of getting in touch with individuals within the rather specific field of our research was by asking previous respondents for recommendations or contacts. Additionally, we reached out to acquaintances in the drug development/medicine industry in order to get in touch with possible interview respondents for our research. This snowball sampling method for finding respondents could potentially lead to bias and a selection of respondents with similar backgrounds (Emerson, 2015). Additionally we reached out to acquaintances in the drug development/medicine industry in order to get in touch with possible interview respondents for our research.

Table 3.1: Respondent and Interview information

Respondent	Role	Organization	Country	Length (min)	Date	Interview Type	Appendix
R1	CEO of post approval AI service, AI advisor in drug development	Company A, Company B	Sweden	26:04	4-19-24	Video & Audio	C
R2	Medicine student & medicine researcher	Institution A	Sweden	22:54	4-19-24	Video & Audio	D
R3	Software developer in pharmaceuticals	Company C	Sweden	21:01	4-23-24	Video & Audio	E
R4	AI in Medicine researcher and author	Institution A	Sweden	N/A	4-19-24	Message (Linkedin)	F
R5	IT R&D in pharmaceuticals	Company D	Sweden	25:33	5-8-24	Video & Audio	G

3.3.3 Interview Guide

Semi-structured Interviews

It was decided that the semi-structured interview method would work well for this study. While there were a few specified subjects that needed to be covered, we also wanted to learn about the interviewees' individual perspectives and encourage them to speak candidly. When capturing the interviewee's perspective is more important than gathering basic information about a subject, the semi-structured approach is a particularly good technique (Adeoye-Olatunde & Olenik, 2021).

The questionnaire was designed to ask open questions but within the topic of research, covering both our theory (XAI) as well as the relevant points of EU ALTAI and AI Act. Supplementary follow-up questions were incorporated to ensure exploration of the specific terms and subjects relevant to our study objectives. Drawing inspiration from Weidener and Fischer's (2024) semi-structured interview questionnaire on ethical AI in medicine, we utilized their questions as a foundational framework. Subsequently, we adapted and refined the questions to align with our research inquiry and theoretical framework. The interview questions employed in our semi-structured interviews can be found in Appendix A.

Interview phases

The interview will be structured in phases. Each phase will have a focus, but we do want to let the interviewees express themselves freely. The phases of the interview are not noticeable during the interview but help us to ensure that we are covering all important aspects of our study. The phases were also used as a guidance in creating the interview questionnaire.

Phase 1: Introduction (with focus on the research) and ethical review

We will introduce our thesis, the research we are conducting and what we hope to gain from the interview. There are also several ethical aspects related to the privacy of the respondents that we will go through, and the respondent will need to approve of these in order to go forward with the rest of the interview. There will also be an introduction to the terminology used in the study.

Phase 2: Background (with focus on the respondent)

The respondent will be asked about their background related to the topic of the interview, and their current work/title in the field. This part of the interview is to gain further insight in the respondents´ credibility. It will also function as a short phase of the interviewers and the respondent getting to know each other.

Phase 3: Exploring Ethical Challenges

The third phase will contain the first interview questions focused on gathering data for our research. The respondent will be asked about their perspective on the topic of ethical challenges with using AI in pharmaceutical drug development.

Phase 4: Addressing Ethical Implications

The respondent will be asked about his or her view on the accountability and transparency related to AI in drug development. The questions will focus on the possible ways of addressing these concerns, as well as the probable or existing obstacles when trying to solve such challenges. In this part of the interview, we hope to gain insight on the first part of our research question, how AI models impact transparency and accountability in drug development processes. This phase is very important as it connects to the theory used in our study.

Phase 5: Exploring Opportunities and recommendations

The final questions will focus on the respondent´s perspective on the benefits that would arise from addressing the concerns and solving the ethical challenges brought up in the interview. The respondents will also be asked about their personal recommendations or beliefs when it comes to solving the ethical challenges with AI in drug development processes. In this phase we hope to gain insight in the second part of our research question, strategies which can be used to mitigate the impact that AI models have on drug development processes.

There is also a possibility that we will get to hear different attitudes towards AI in drug development processes in this phase, which would add an interesting aspect to our research.

Phase 6: Wrap-up and finalization

The final phase of the interview will start with the respondent getting to add any additional information if they wish to do so. Thereafter we will finalize the interview by thanking the respondent for their participation and terminating the interview.

3.3.4 Ethical Considerations

When undertaking qualitative research, it's crucial to recognize and address the ethical considerations involved in engaging external participants. All individuals involved should be fully briefed on their ethical entitlements as participants, as well as how their interview data will be utilized within the study. Additionally, participants should be provided with details regarding the study's objectives and potential benefits. They should also be made aware of their right to withdraw from the study at any point if they so desire (Oates, 2006).

Throughout our interview process, we ensured that ethical information pertinent to the respondents was integrated into an informed consent document (see Appendix B). This document was distributed to participants at the time of interview scheduling. It's important to note that this document was not designed as a legally binding contract for participation; rather, its purpose was to reassure participants of their rights (Oates, 2006). Prior to their interviews, participants were required to sign this consent form as confirmation of their understanding and agreement.

For the purpose of transcription of the interviews, we asked the participants if they approved of us recording the interview. All participants agreed to this, which provided good conditions in order to transcribe the interviews.

3.4 Data Analysis

Thematic analysis is a prominent method of analysis in qualitative research. The emphasis is on identifying, assessing, and interpreting meaning patterns (also known as "themes") in qualitative data (Braun & Clarke, 2006). Thematic analysis is used in qualitative research to identify themes or relevant patterns in data (Daly et al., 1998). This strategy can highlight the data set's organization, detailed description, and theoretically justified interpretation of its relevance (Braun & Clarke, 2006).

To transcribe the data gathered from interviews, we employed Whisper, a software tool specialized for this purpose. This allowed us to solely focus on the interviews and collect relevant data, streamlining the transcription process and ensuring accuracy while minimizing distractions. Whisper operates by converting audio recordings into written text through advanced speech recognition algorithms. It offers features tailored specifically for transcription tasks, such as customizable settings for different accents and speech patterns, enhancing accuracy. Additionally, Whisper's interface facilitates easy navigation and editing of transcribed text, allowing users to quickly correct any errors and refine the transcript as needed. Its efficiency enabled us to expedite the transcription process, freeing up valuable time for analysis and interpretation of the data.

When working with the data and applying the thematic analysis we used a two-sided approach. We used ChatGPT to help us with our analysis while we also kept Bruan & Clarke's (2006) six-step approach for thematic analysis in mind; Familiarizing yourself with the data, generating initial codes, searching for themes, reviewing themes, defining, and naming themes and producing the report.

Initial Code/Foundation Process

According to Kruikow (2024), the next crucial step in our thematic analysis process involves working with the transcript data through various stages. As Kruikow explained (2024) our thematic approach comprised three overall steps: initial coding, focused coding, and creating themes. Initially, we laid the groundwork by breaking down the data and generating preliminary codes. Then, we conducted focused coding to refine and organize these codes, identifying recurring patterns. Finally, we synthesized the coded data into overarching themes, providing an understanding of our research findings.

We employed ChatGPT for our thematic analysis, ensuring a careful procedure to guarantee validity and transparency. The following steps were taken:

1. Establishing Transparency and Control: Recognizing the importance of maintaining transparency and avoiding shortcuts, we implemented continuous monitoring to uphold trustworthiness and validity while controlling for bias.

(Kruikow, 2024).

2. Utilizing Microsoft Word for Transparency: In our analysis, we integrated Microsoft Word alongside ChatGPT to centralize all collected data, enhancing transparency throughout the process.

(Kruikow, 2024).

Laying Groundwork

This phase is critical in the thematic analysis process because it requires the production of a comprehensive list of descriptive and detailed codes (Kriukow, 2024). These codes help to have a thorough knowledge of the data (Kriukow, 2024). Kruikow, (2024), emphasizes that it is vital to note that this phase poses issues when utilizing ChatGPT because it may propose shortcuts, perhaps resulting in data loss. Avoiding premature analysis is critical for ethical transparency and limiting bias to assure the validity of findings, and comprehensive prompts are required to accurately direct ChatGPT while providing context and avoiding confusion (Kriukow, 2024).

As per Kruikow (2024), when coding the transcript, providing detailed prompts is crucial for effective guidance in the process. These prompts should strike a balance, offering enough information to explain the task at hand and provide context for ChatGPT to understand the requirements without overwhelming it (see Appendix H). Our goal was to break down the transcript into smaller units of analysis for coding. The codes we assign must be descriptive, capturing the content accurately rather than being abstract (Kriukow, 2024). Including examples well-crafted in the prompt and throughout the process helps maintain clarity. Moreover, Kruikow (2024) discloses that it's essential to be clear that you want to create a comprehensive list of codes, ensuring that all aspects are coded to prevent ambiguity. Without this thorough approach, there's a risk of codes resembling abstract themes rather than specific content.

Additionally, each code was accompanied by a full list of quotes connected to the participants' interviews to facilitate coherence in our analysis and make it easier to refine interpretations later on (see Appendix I). We got these quotes by pasting a second prompt commanding chatGPT to assign quotes from the participants to each code (Kriukow, 2024). Without these quotes, codes lack context and utility. They ensure transparency, accurate reporting, and effective monitoring of the analysis process. In summary, compiling these quotes is vital for maintaining validity and coherence in our thematic analysis (Kriukow, 2024).

To streamline our organization and analysis, we created a separate document in Microsoft Word or another external platform for each interview participant as Kruikow (2024) suggests. In these documents, we pasted the list of codes associated with the quotes from each participant's interview. For clarity, we considered color-coding each participant's codes uniquely.

Additionally, we created another document, labeling it "Focused Coding," and pasted the output of the second prompt, containing only the codes without the quotes (Kriukow, 2024). This document served as our repository for focused coding. Following Kruikow's (2024) approach to thematic analysis we had to constantly ensure that we maintained consistency by using the same color scheme as the participant documents for each code. By doing so, we could easily differentiate between the focused codes associated with each participant, facilitating clearer analysis and interpretation (Kruikow, 2024).

Focused coding

As Kruikow (2024) suggests, focused coding is a manual process to ensure transparency and validity. Instead of themes, we group codes into categories to better understand and organize them. Through moving, copying, cutting, and pasting, we created groups with general trends and patterns, comprising codes from different sources or datasets (Kruikow, 2024).

However, this manual process can sometimes lead to confusion about the code's content (Kruikow, 2024). To address this, we referred to the original source files containing the initial codes from individual participants (Kruikow, 2024). The color-coding system helps us locate the correct file and quote, aiding in recollection (Kruikow, 2024).

Another challenge was the occurrence of duplicates within groups. While these duplicates may vary slightly in wording, they essentially convey the same meaning. To manage this, we created a new description and pasted the original duplicates as comments, maintaining their color (Kruikow, 2024). This practice ensures transparency and allows us to maintain control over the process (Kruikow, 2024).

Creating themes

Drawing from Kruikow's (2024) suggestions, creating themes is preferably a manual process where we examine the focused codes to organize them into coherent themes or topics that effectively communicate our understanding of the data to the reader. Understanding the data is crucial in this step. While we could use ChatGPT to assist in organizing the codes into themes that address the research question or prompt, the final step involved reorganizing the codes to construct a narrative that effectively communicates the answers to the research question (Kruikow, 2024). This manual approach ensured that the themes accurately represented the data and conveyed the intended message to the reader (Kruikow, 2024).

After initial coding, codes were organized into potential themes that accurately reflected the collected data. Each theme underwent a careful examination to ensure cohesion and distinct representation of the dataset aspects, followed by a review to confirm their relevance and clarity in relation to the research questions. Adjustments were made to better capture the essence of the insights gathered. Finally, each theme was clearly defined and named, ensuring they were ready for presentation and clearly reflected the depth and nuance of the qualitative data (Braun & Clarke, 2006).

Table 3.2: Final table of themes

Theme	Code	Colour
Ethical AI and Transparency	EA	🟥
Regulatory Alignment and Challenges	RAC	🟦
Stakeholder Perspectives and Industry Challenges	SPIS	🟩
Balancing Innovation and Ethical Standards	BIES	🟪
Integration of AI in Clinical and Pharmaceutical Ethics	IACPE	🟫

To clarify and simplify the later analysis of the transcripts when assigning themes, a thematic color-coding system was employed. Each theme identified in the study was assigned a specific color, as detailed in Table 2 under Section 3.4 'Data Analysis.' This method enables the rapid detection and categorization of linked textual snippets and quotes, resulting in a more efficient analytical procedure. When a text snippet could potentially be associated with two different themes, we have opted to color the text in one of the theme colors and the background in the other theme color. In cases where a snippet could apply to three different themes, an assessment is made to determine which two themes are most suitable for that snippet. This approach allows for easy visual differentiation of thematic elements and helps track their recurrence and distribution throughout multiple interviews. While this technique improves access and depth of knowledge by using visual clues to reveal complex patterns and relationships within the data, it could be considered a limitation, as it may require subjective decision-making in selecting the most relevant themes for mixed-content snippets.

3.5 Validity and Reliability

Nowadays, it's critical that qualitative researchers approach the design of their studies with rigor and imitate the scientific process by aiming on empirical groundness, generalizability, and bias minimization. There are two ways that researchers can make sure that qualitative studies are of high quality: by making them more reliable and valid. Increasing the reliability and validity of qualitative studies require the use of research techniques that guarantee accurate data recording and empirical, logical, and replicable data interpretation (Franklin & Ballan, 2001).

While interpretations of reliability have varied over time, its core essence remains constant and Joppe (2000) captures that and defines reliability as:

> *"...The extent to which results are consistent over time and an accurate representation of the total population under study is referred to as reliability and if the results of a study can be reproduced under a similar methodology, then the research instrument is considered to be reliable."* (Joppe, 2000, p.1)

While the concept of "reliability" is often associated with quantitative research, it is also commonly applied to other types of study (Golafshani, 2015). It can be explained as follows: when reliability is a concept that is used to evaluate quality in a quantitative study, quality is evaluated with the "purpose of explaining," whereas in a qualitative study, quality is evaluated with the "purpose of generating understanding." (Stenbacka, 2001, p. 551).

Many different phrases are used in qualitative research to describe the idea of validity (Golafshani, 2015). While some qualitative researchers maintain that the term validity does not apply to their type of study, they have also acknowledged the necessity of a qualifying check or measure (Golafshani, 2015), According to Wainer and Braun, (2015) in qualitative research, validity refers to how "appropriate" the methods, procedures, and data are. Whether the design is valid for the methodology, whether the sampling and data analysis are appropriate, whether the research question is valid for the intended outcome, and whether the results and conclusions are valid for the sample and context (Wainer & Braun, 1988).

To maintain the reliability and validity of our qualitative research, we have purposefully used a number of methods. As we conducted our research, we kept Stenbacka's (2001) distinction in mind. We focused on understanding rather than just explaining. This influenced our choice of methods, like using semi-structured interviews and selecting deeply involved participants. By prioritizing understanding, we aimed to uncover the depth and complexity of our topic, contributing to qualitative research's broader goals. We sought to guarantee the robustness of our findings, which we defined as the stability of outcomes across time and their correctness in representing the entire population under study as Joppe (2000) stated. In relation to that and in addition to conducting interviews, we used a variety of techniques, including participant observation and document analysis, to double-check our findings and improve dependability. We tried to improve our study's reliability and repeatability by using a variety of sources and methodologies, in keeping with the tenets of high-quality qualitative research. Examples of that would be only using academic journals, published and peer reviewed papers.

Furthermore, we upheld Wainer and Braun´s (2005) definition of validity by incorporating peer-reviewed past research and articles. Beginning with a well-defined research question, we ensured alignment with established literature and theoretical frameworks. We took great care in our interview selection process, choosing people who were well-established in the field, in keeping with Wainer and Braun´s emphasis on appropriateness. This together added to the overall credibility of our research project.

4 Empirical Findings

In this chapter we present our empirical findings from data collection and analysis. Findings are structured based on the thematic analysis which was created based on interview transcriptions.

4.1 Theme 1: Ethical AI and Transparency

In the realm of pharmaceutical development, the integration of Artificial Intelligence (AI) brings to the forefront critical ethical concerns, particularly around transparency and explainability. As outlined in our empirical findings, the theme "Ethical AI and Transparency" captures the industry's imperative to align AI technologies with ethical standards that promote trust and accountability.

Respondent 2 presents an argument on the interplay between AI transparency and ethics. They highlight the fundamental link between explainability and moral AI practices, suggesting that ethical validation cannot be achieved just through traceability. They emphasize,

> "*I believe that explainability and ethics must go hand in hand. It's not the case that just because something is fully traceable, it is automatically ethical."* (R2:20).

This statement underscores the necessity for AI systems to not only track but also clearly justify their processes and decisions, ensuring that ethical considerations are included into the creation and use of AI. Building on the importance of comprehensible AI, Respondent 1 raises concerns about the feasibility of peer reviewing AI-driven research when the underlying algorithms remain opaque. They highlight a significant barrier to scientific validation:

> *"It is definitely a significant ethical challenge because there needs to be traceability... If you then consider what you call Black Box AI. I don't even know what algorithms are behind it. How can it then be peer reviewed in the next step" (R1:8).*

In agreement with this, Respondent 5 demonstrates the critical function that explainability plays in guaranteeing user trust and ethical recognition:

> *"Say I am looking for information about holiday entitlements in Sweden. AI can find that information and say that in Sweden, you have 35 days of holiday. Can I directly trust this information? Perhaps not. But if the AI-generated answer also includes a reference to the document where the answer was found, then I can go there and validate it"* (R5:8).

This highlights the necessity of explainability in AI systems, which is essential for verifying results and building trust, especially in critical fields like drug development.

Further discussing the risks associated with opaque AI, Respondent 1 also touches on the potential public health risks when decisions made by non-transparent AI systems affect millions, particularly in critical areas such as pandemic response. Highlighting the dire consequences of deploying unverifiable AI technologies at scale, they caution:

> *"And I believe, for the sake of transparency, one must show which models are behind it. Because this can indeed affect millions of people" (R1:10),*

Respondent 3 expands on the practical implications of explainability, noting its crucial role in ensuring patient safety and effective drug development pathways. They discuss the importance of having clear and interpretable models:

> *"Safety is paramount; we don't want patients to experience unnecessary side effects... Explainability is crucial, especially considering the various stages involved in drug development" (R3:6).*

This insight stresses that each development stage must be transparent to predict outcomes accurately and navigate the path to market successfully.

Adding to the dialogue, Respondent 4 reflects on the broader implications of AI within the drug development sector, emphasizing the need for an ethical framework that includes transparency, fairness, and patient safety. They assert,

> *"Ethical implications of AI in drug development are multifaceted, but a central concern revolves around transparency, fairness, and patient safety. Ensuring that AI systems can provide explanations for their decisions is crucial for transparency, accountability, and maintaining trust between AI systems and the humans they interact with"* (R4:5).

This comprehensive view encapsulates the necessity for ethical guidelines that govern AI usage to ensure equitable outcomes and maintain the integrity of medical advancements.

In promoting a cooperative strategy between AI and human validation, Respondent 5 emphasizes how important it is to lower the risks and ethical questions surrounding using a "black box" solution:

> *"In this way, we not only decrease the ethical issues but also the risk associated with using a 'black box' solution, as AI cannot act autonomously if we do not have a full understanding of its function."* (R5:12).

This viewpoint underscores the inherent limitations of AI autonomy when its operations are opaque, stressing the importance of transparency and comprehension to address ethical and operational risks. Furthermore, Respondent 5 highlights the need to apply basic techniques to the field of AI-based systems for scientific application:

"*This also applies to AI-based systems. To use AI in a scientific manner, one needs to transfer the fundamental methodologies into the AI realm."* (R5:14).

This claim emphasizes how crucial it is to incorporate accepted scientific methods into AI frameworks in order to guarantee their level of trustworthy use in R&D procedures. Respondent 5 also emphasizes the vital role explainability has in AI throughout this integration process:

"In this process, explainability becomes important in AI, as it helps us to understand and validate our results." (R5:14).

In summary, Respondent 5 underscores the importance of transparency, comprehension, and the integration of established scientific methods into AI frameworks to mitigate risks and ethical concerns associated with 'black box' solutions, emphasizing the vital role of explainability in ensuring the reliability and trustworthiness of AI-driven insights.

Conclusively, the collective insights from our respondents show the widespread concern for ethical standards in AI applications within drug development. By underscoring the essential for transparency and explainability, the discussion aligns with regulatory expectations and ethical norms, aiming to foster an AI-integrated future in healthcare that is both innovative and ethically sound.

4.2 Theme 2: Regulatory Alignment and Challenges

The talk on "Regulatory Alignment and Challenges" focuses on the dynamic interaction between the need for strong regulatory frameworks and the development of artificial intelligence (AI) technology in medicines. The many challenges and requirements that the pharmaceutical business must fulfill when integrating AI are highlighted by respondent insights.

Respondent 3 expresses what regulatory bodies hope to see in terms of transparency and a solid regulatory future:

"Regulatory authorities also expect this level of transparency, so I don't foresee it changing" (R1:8).

This expectation sets a baseline for how AI applications should be managed, emphasizing that transparency isn't just a best practice but a continuing regulatory demand.

Furthering the conversation, Respondent 2 stresses the importance of regulatory specifics that ensure AI models meet certain standards, especially in sensitive industries:

"There surely needs to be some kind of regulations where the AI model is sufficiently covered. Like showing what it has been trained on, especially when it concerns the pharmaceutical industry." (R2:18).

This statement emphasizes the necessity of well designed laws that are not just broad but also specifically tailored to handle the particular difficulties presented by AI in healthcare, guaranteeing both safety and effectiveness.

Adding to this, Respondent 5 brings up additional complexities related to 'black box' AI systems, highlighting issues around accountability and the evolving nature of AI models:

> “*There must always be a legally responsible person behind it all”* (R5:12). They continue “*The biggest problems with 'black box' solutions in terms of AI or models have more to do with intellectual property rights”* (R5:12).

Further, Respondent 5 mentions,

> *“But there's also something known as 'model drift', where the model changes over time as it learns more. Eventually, it can start detecting incorrectly”* (R5:20).

This insight underscores the importance of continual monitoring and adaptation of regulatory frameworks to keep pace with the technological advancements and inherent unpredicability of AI systems.

Respondent 1 reflects on the European Union’s efforts to shape the regulatory landscape, though with mixed feelings about the current robustness of these regulations:

> *"Now the EU is making an attempt to regulate the AI market. It's somewhat a fluffy document but it still provides directions."* (R1:12).

This comment suggests a recognition of the steps being taken, yet also a call for more concrete and detailed regulatory frameworks specifically designed for the pharmaceutical and medical device industries.

When taken as a whole, these observations paint a picture of a dynamic regulatory environment where there is broad agreement about the need for laws to be more enforceable, transparent, and flexible in order to keep up with new developments in technology. In order to guarantee that AI technologies be applied in a way that is ethical, safe, and advantageous to public health, the respondents as a whole support a regulation strategy that is both proactive and reactive.

4.3 Theme 3: Stakeholder Perspectives and Industry Challenges

The topic of "Stakeholder Perspectives and Industry Challenges" captures the differing opinions of important players in the sector about the application of AI to medication development and healthcare. The views of many responders illustrate the complexities involved in integrating AI into delicate domains like patient care and clinical decision-making, emphasizing the necessity of cooperation, openness, and trust.

Respondent 3 articulates the significance of explainability in fostering trust among stakeholders, particularly healthcare professionals who rely on AI tools. They note,

> *"So, when it comes to explanations, it's not only about the technology but also about convincing the stakeholders, such as physicians, to trust the process"* (R3:13)

The acceptance and adoption of AI in clinical settings hinge significantly on these stakeholders understanding and trusting how AI decisions are made. Respondent 3 further illustrates these complexities involved in integrating AI into patient care, particularly the challenges of convincing physicians to trust AI recommendations:

> *"However, it was challenging to convince physicians to trust these doses. We had to conduct numerous presentations, not only providing the dose but also presenting graphics to explain why it might be higher or lower than what they typically prescribe"* (R3:13).

This necessity for detailed explanations and visual aids underscores the importance of transparency in fostering trust within the medical community. Similarly, Respondent 1 expresses concerns about the acceptance of AI without full explanatory transparency, highlighting the hesitation within the medical field to endorse analyses whose methodologies are not openly disclosed:

> *"So personally, I would find it very difficult to stand up and say that here we have a great analysis but I can't tell you how it was made."* (R1:10).

Both respondents point to a common industry challenge: the need for AI processes to be understandable and verifiable to gain clinical acceptance and trust.

Expanding on the need for broader engagement, Respondent 4 discusses the importance of collaboration across various stakeholders to address ethical considerations and align AI development with societal values:

> "*Engaging with healthcare professionals, regulators, patients, and other stakeholders to solicit feedback, address concerns, and co-create solutions that prioritize ethical considerations and align with societal values"* (R4:10).

This engagement is crucial not only for the adoption of AI but also for ensuring its alignment with ethical standards and regulatory requirements. Respondent 4 also highlights AI’s potential to tackle significant global health challenges, providing a vision for the positive impact of AI on global healthcare outcomes:

> *"The application of AI in addressing global health challenges, such as antimicrobial resistance and rare diseases, presents exciting opportunities for leveraging technology to improve healthcare outcomes worldwide"* (R4:16).

Together, these insights emphasize the necessity of a multifaceted approach involving clear communication, ethical considerations, and collaborative efforts to navigate the challenges and harness the benefits of AI in healthcare.

These perspectives underline a shared understanding among stakeholders: the successful integration of AI into healthcare depends on transparent, comprehensible AI applications supported by ongoing dialogue and cooperation among all parties involved. This approach ensures that AI advancements are not only technologically sound but also ethically responsible and widely accepted within the medical community and beyond.

4.4 Theme 4: Balancing Innovation and Ethical Standards

The theme "Balancing Innovation and Ethical Standards" balances the need to follow ethical standards with the quickening pace of artificial intelligence (AI) developments in the healthcare industry.

Respondent 4 sheds light on the transformative capabilities of AI in the pharmaceutical industry, emphasizing the role of cutting-edge technologies in enhancing drug development processes:

> *"Advances in reinforcement learning and deep learning techniques are opening up new possibilities for optimizing drug design, predicting drug-target interactions, and accelerating the drug discovery pipeline"* (R4:16).

This optimistic view of AI's potential is balanced with a caution about the ethical implications, stressing the need to

> "*address these concerns that require a balance between innovation and ethical oversight to uphold patient trust and safety in medical AI applications"* (R4:8).

Adding to this perspective, Respondent 5 highlights the accelerating capabilities of AI:

> *“'AI can then significantly accelerate the process. For example, if we need to test 5,000 or 10,000 different variants of a molecule and how it can be modified, AI makes it much easier for us to select a fraction that it suggests as reasonable alternatives”* (R5:18).

This illustrates the practical efficiency brought by AI while maintaining a focus on ethical considerations.

Building on this point, Respondent 1 raises concerns about the autonomy of AI systems and the critical need for human oversight to ensure quality and reliability:

> *"Do we dare to let AI alone summarize this? Or must we also actually have some skilled medical people read through all this and quality assure what AI has done?"* (R1:6)

Respondent 1 advocates for a collaborative approach where AI does not replace but rather enhances human capabilities, suggesting a synergy that leverages AI to improve outcomes while retaining human judgment to ensure ethical integrity by emphasizing that:

> *"I believe that it must be used in combination a lot with human intelligence and not become any form of slave to AI"* (R1:17).

Respondent 5 reinforces this idea, stating,

> *“Despite AI helping us to accelerate the whole process, it does not replace humans in any of the steps. Instead, it reduces the effort required to get through all the steps”* (R5:18),

which aligns with the necessity of human involvement in critical decision-making processes.

Respondent 2 highlights the challenge of inherent biases in AI systems, pointing out the critical need for transparency and accountability in how these systems are trained: *"*

> *And there's a lot, a big bias is, what information does this have from before"* (R2:9)?

This comment highlights the crucial need that AI systems are developed and run in a way that recognizes and reduces any potential biases, especially when these systems are applied in a variety of complex and diverse domains like healthcare. Furthermore, Respondent 2 emphasizes the importance of rigorous ethical scrutiny in AI applications:

> *"It must be well proven that it has good judgment, in my opinion. There should be strict requirements to compromise with the other ethical aspect."* (R2:20).

This insistence on proven reliability and ethical compliance highlights the stakes involved in deploying AI within contexts where decisions have profound impacts on human lives.

Further expanding on the context of AI's integration, Respondent 3 discusses the ongoing evolution towards full AI integration, noting the current and potential contributions of AI to their work:

> *"Regarding AI, I must mention that while we're not fully integrating it at the moment, we're certainly moving in that direction.”* (R3:4).

They detail the practical applications of AI, such as ‘covariate searching’, which involves analyzing patient characteristics that influence drug efficacy. This realization emphasizes how AI can be used to improve comprehension and treatment effectiveness in real-world settings—if ethical issues are taken into account throughout system implementation.

4.5 Theme 5: Integration of AI in Clinical and Pharmaceutical Ethics

AI integration in pharmaceutical and clinical settings is a hot topic in ethical discourse, especially when it comes to striking a balance between AI's potential and the strict ethical norms required in the healthcare industry. The perspectives offered by respondents with varying backgrounds shed light on the challenges associated with applying AI in a way that is both morally just and beneficial to medicine.

Respondent 1 and Respondent 4 highlight the collaborative potential between AI and human expertise in enhancing diagnostic accuracy. As Respondent 1 notes, the combined efforts of AI and skilled radiologists in breast cancer screenings

> *"For example, in breast cancer screening, AI can make a fantastic interpretation of a mammogram. And so can a skilled radiologist. But the two together produce even better quality.*" (R1:17).

showcase how AI can augment human capabilities rather than replace them. Complementing this perspective, Respondent 5 emphasizes the supportive role of AI in the process, stating,

> “*Although AI helps us accelerate the entire process, it does not replace humans in any of the steps. Instead, it reduces the effort required to get through all the steps*.” (R5:18).

This reinforces the idea that AI is a tool that enhances human decision-making and efficiency without supplanting the essential human elements needed in medical diagnostics. Adding another perspective, Respondent 4 underscores the necessity for robust ethical oversight when deploying AI technologies:

> *"The ethical challenges with AI certainly affect my work... Understanding how AI models make predictions or recommendations in drug development is crucial for ensuring the safety and efficacy of pharmaceutical products"* (R4:10).

These discussions emphasize the need for a synergy between technological advancement and ethical vigilance to ensure optimal patient outcomes.

Further detailing the integration of AI in pharmaceutical processes, Respondent 3 sheds light on the specific applications of AI in drug development, particularly in the management of complex patient data:

> *"We're certainly moving in that direction. Given that modeling is our focus, it's natural that AI is becoming increasingly important."* (R3:4).

They elaborate on how AI facilitates the handling of covariates in clinical trials, enhancing the precision and personalization of patient care. This operational benefit is tied closely to the ethical considerations discussed by Respondent 2, who praises the potential of AI to deliver crucial medical recommendations:

> *"The greatest benefit would be that it can give recommendations. In healthcare, recommendations are crucial, and if AI can provide them, it would be a major success." (R2:22).*

Further insights into the ethical integration of AI in clinical settings come from Respondent 3, who provides a practical perspective on the operational aspects of AI in drug development. They note the specific tasks AI assists with:

> *"So we basically develop models based on clinical data to address various questions that can arise during drug development. These questions could range from whether to continue with a particular drug to selecting another dose or determining the design for the next stage of drug development studies."* (R3:4).

The enhancement of decision-making processes through AI highlights its potential to significantly influence patient outcomes positively. They further elaborate on the complexity of integrating AI, particularly in the realms of pharmacometrics and patient safety,

> *"Firstly, because pharmacometrics is quite a niche field, maybe involving only a few hundred people in Europe. Therefore, we must prioritize safety and efficacy"* (R3:6).

Respondent 4 delves deeper into the ethical considerations necessary for integrating AI within pharmaceutical frameworks, stressing the importance of maintaining high standards of scientific rigor and transparency:

> *"In drug development, decisions based on AI models can have significant scientific implications. Transparent and traceable AI models enable researchers to validate findings, reproduce experiments, and ensure the reliability of results"* (R4:22).

They highlight how ethical frameworks not only guide the development of AI technologies but also ensure their alignment with industry standards and societal expectations, thus fostering a trust-based relationship between technology providers and healthcare recipients. Building on this, Respondent 5 articulates the importance of a deeper understanding of AI applications, stating,

> *"By doing this, one can increase trust in the results obtained. Understanding is key; not just for the specific experiment or the system's area of use but also for future knowledge development.*

> *Scientific progress builds on past knowledge and continuous improvements of models to take them to the next level.*" (R5:14).

This statement reinforces the vital role of comprehensive understanding and ongoing improvement in AI technologies, which is necessary for advancing medical science and maintaining ethical integrity.

Respondent 5 continues by addressing a more societal issue that can come with incorporating AI in the industry by questioning,

> *"Will it lead to a greater part of the workforce being replaced, and so on? Is it the right thing to do from a societal perspective?"* (R5:10).

The same respondent further goes on to address another ethical issue with the specific industry, namely,

> *"If you're working with access to large amounts of health data and using AI to draw conclusions, such as a patient having a very high likelihood of developing cancer, should you then inform the person and take preventive measures if you have that information available?"*" (R5:10).

This question raises critical ethical considerations about patient rights and the responsibilities of healthcare providers, underscoring the complex interplay between innovation and ethical standards in AI applications.

5 Discussion

In this chapter, we explore the European Union's regulatory frameworks AI Act and ALTAI, to fully understand the EU requirements discussed in this study. Additionally, we analyze the empirical evidence, comparing it with previous research on how the drug development industry views the use of artificial intelligence in drug development. Finally, by synthesizing research findings, academic literature, and regulatory frameworks, we address the research question: How does the drug development industry's perspective on the ethical concerns posed by Black Box models correspond with EU requirements surrounding AI use?

5.1 Foundation of Theoretical and Regulatory Perspectives

In this research, the two main regulatory frameworks analyzed are from the European Union. One serves as a self-assessment tool while the other will soon function as a legal framework, setting binding regulations for AI applications. Both frameworks address the issues of 'Black Box' models in AI and the need for explainability. Additionally, the EU is actively engaging in research on Explainable Artificial Intelligence (XAI) to reduce reliance on AI systems that lack complete transparency.

5.1.1 Requirements for Black Boxes and XAI common for ALTAI and AI Act

Both documents emphasize the importance of ethics and transparency in AI outputs, advocating that XAI should be the principle to aspire for in all AI use, however both frameworks also accept some level of Black Box AI. According to the ALTAI framework by the European Commission (2020), the level of explainability needed in an AI system is determined by the effects a failed output would have on a human life (European Commission, 2020). In the case of AI in drug development, where a failed drug could be vital, we consider the need for explainability to be high, thus the acceptance for black box models within this field to be rather low. Similarly, Panigutti et al. (2023) interpret the AI Act framework to encourage that high-risk AI should use XAI as much as possible, but it does not entirely forbid the use of black box AI (Panigutti et al. 2023; European Parliament and Council of the European Union, 2024). This also indicates low acceptance for Black Box AI since medical devices are ranked as high-risk AI systems in AI Act (European Parliament, 2023). The EU, based on both frameworks and reflections of these made by Panigutti et al (2023) and the EMA (2023), advocates for a human-in-the-loop approach to AI, and argues that a human-centric approach is the best handling of complex AI models.

The two frameworks together represent the EU requirements on ethical use of black box models and will function as the foundation for comparison towards empirical findings and previous research.

5.2 Perspectives of the Drug Development Industry

This section explores the drug discovery industry's viewpoints on integrating AI into drug development, along with the associated ethical considerations. It includes empirical data and insights drawn from healthcare leaders and organizations, based on prior research.

5.2.1 Ethical AI and Transparency

The description of ethical AI is described as a set of values and principles used to guide moral conduct in the use of AI (Leslie, 2019). Within the drug development industry there seems to be consensus that the ethical aspect of AI holds a very strong importance due to patient safety and health. Our findings describe patient safety to be paramount, stating that explainability is crucial in ensuring exclusion of patient side effects. Patient safety and explainability are, according to our findings, central concerns of AI use. This is an attitude to AI use which is shared with healthcare leaders in the BMC Health Services Research (Petersson et al., 2022) who state the particular importance of ethics in AI when it affects human health. The sentiment is further emphasized by Iniesta (2024) as well as Durán & Jongsma (2021) who highlight the necessity in explainability when AI is used in a medical context, which our findings address further with a statement that transparency in the drug development context could have effects on millions of people.

We can see that the drug development industry has a common understanding that patient health and safety is at the highest priority when integrating AI in drug development, leading to them placing high value in the explainability of AI systems. Our perception of these results are that the drug development sector is characterized by the responsibility they feel in creating powerful drugs which are meant to help people, but can also cause harm if not handled responsibly.

Traceability and validity of results was also a topic brought up on the theme of ethical AI and transparency. Our findings indicate that it is a problematic situation when not knowing what algorithms are behind an output, which can lead to issues in the following step of the results being peer reviewed. Several of our respondents mention the importance of explainability in AI models to allow validation of results as well as enabling trust between the user and the system. These findings can confirm what healthcare leaders in the previously mentioned BMC study (Petersson et al., 2022) agree on, that explainability concerns with black box models are significant ethical issues.

5.2.2 Regulatory Alignment and Challenges

Our interviews did not ask about regulatory or requirement alignments, or a need for such a thing, however, this topic came up in several interviews. Our findings show that the respondents consider there to be a need for regulations regarding the use of AI, showing that regulations regarding explainability are expected and necessary as this can give direction. One example brought up by Respondent 4 is that there need to be regulations on things such as what data an AI has been trained on. Healthcare leaders in the BMC study (Petersson et al., 2022) express their critique towards AI regulations not keeping up to date with AI advancements. Furthermore, Weidener & Fischer’s study (2024) shows that developers of medical AI applications find it hard to know where to draw the line between innovation and ethics, implying a need for clearer direction. Although these statements are not identical, we see a connection in the perspective that AI regulations are needed in one way or another, and that they need to be up to date.

The discussion of the topicality of AI regulations can according to the BMC study result in stifling the progress of AI within healthcare (Petersson et al., 2022). Furlong (2023) presents similar opinions from Hopkins and Weatherhall who are both eminent figures in the drug development industry. They agree that overregulating the area of AI in drug development is to be avoided since there is already much existing regulation surrounding drug development and medicine, also stating that there needs to be a more progressive approach to regulating AI use in drug development. Medtech Europe (2024) also addresses the issue of overlapping regulations, arguing that this can lead to uncertainty in development and delivery of new AI innovation.

Our findings mainly show the perspective of the need for requirements and regulations, while previous research addresses the problems with outdated regulations and uncertainty in overlapping regulations. Although the statements on this theme do vary a bit, we see a connection in perspectives on the need for regulations, even if it is argued that they need to be both more up to date, progressive and intertwined with overlapping regulations.

5.2.3 Stakeholder Perspectives and Industry Challenges

Trust emerged as a recurring theme throughout our interviews. The findings indicate that trust is a complex issue not only in the relationship between users and machines, but also between users and stakeholders. Several respondents provided examples illustrating the challenges of persuading stakeholders or physicians to accept analyses or results derived from AI systems. These challenges are twofold: practically, in convincing others of the validity of results without sufficient explainability, and morally, as one respondent spoke on ethical dilemmas in advocating for AI-generated outcomes which they do not comprehend fully themselves.

Although our initial questions didn't directly address the trust between user and stakeholder perspective, the subject emerged spontaneously during several interviews, particularly in discussions about AI and explainability. This revealed concerns about the moral responsibility associated with making decisions based on systems that aren't fully understood. This topic was not anticipated and is not well-documented in our literature review, suggesting its novelty or under-representation in existing research. Despite this, the recurring nature of these discussions among our respondents underscores the significance of the topic. This identifies a need for further investigation to ascertain the trust between AI users and the stakeholders, and its broader relevance and implications within the industry.

5.2.4 Balancing Innovation and Ethical Standards

Our research highlights a strong consensus on the effectiveness of AI in enhancing work efficiency and significantly speeding up drug development processes. This view is prevalent across the drug development sector and is recognized as a fundamental belief in existing literature. Nonetheless, there are various opinions on how much ethics should control innovation. In our findings, a very ethical attitude was detected. Our analysis reveals a predominantly ethical stance, marked by a cautious approach to potential ethical dilemmas such as bias in drug development. A similar stance is held by Durán and Jongsma (2021) in their suggested approach to black box models. Although they argue that total explainability is not necessary in order for an AI to be trustworthy, they do agree that ethics cannot be deprioritized (Durán and Jongsma, 2021). This position resonates with Weidener & Fischer’s (2024) findings, which emphasize "consequences" as a crucial aspect of their findings. This aspect highlights the imperative for those involved in AI to be cautious of potential biases and discrimination, underlining the essential role of ethics in AI development and application.

While Weidener & Fischer’s (2024) study corroborates the cautious stance towards ethics in AI seen in our findings, however, it also presents a contrasting viewpoint. These developers perceive the ethical considerations around AI not just as guidelines, but as potential obstacles that could hinder the broader progress and evolution of AI technologies.

On the same theme, another topic arose, namely the one of human control and involvement in AI innovation. Our findings show that the expectation is that AI will not work, which the respondents mean is a necessary part of conducting ethical work with AI. Several respondents state that the best result comes from the machine and the human working together, not one without the other. This is a sentiment which is agreed on by EMA (2023), when they state the need for human validation of AI, and that AI is intended to enhance rather than replace human expertise. Also, Durán and Jongsma (2021) match these statements when they argue that physicians need to make mindful decisions on results presented by AI.

The general attitude in the discussion on balancing innovation and AI shows to be that ethics weighs higher than innovation, emphasizing that the best way to balance the two is to not let an AI work alone, but in collaboration with human morals. However, divergent opinions in Weidener & Fischer’s (2024) study, which predominantly reflect developers' perspectives, contrast with our findings, suggesting that their focus might influence their outcomes. Despite exploring potential links between their results and responses from developers in our study, no similarities were found. This discrepancy highlights an intriguing area for future research to explore why these perspectives differ.

5.2.5 Integration of AI in Clinical and Pharmaceutical Ethics

In this theme on integration of AI in specifically the ethical context of pharmaceuticals and drug development, human involvement was again a topic which surfaced many times. To build upon this same subject from the previous section (5.2.4), our findings demonstrate a strong belief that human oversight holds an especially high importance in drug development. Our findings also present an issue which does not regard specifically the drug development industry, but is an ethical issue nonetheless, namely the risk that incorporating AI to an extent where human control is rarely needed would cause loss of work for individuals working in this field, leading to societal issues in the end. However, to judge from the whole of the discussion, this is not a possibility which representatives from the industry believe to be a risk, due to the sensitive nature of the drug development field.

6 Conclusion and Recommendations

In this chapter, we summarize our findings, present answers to the research questions, and discuss the study's limitations. Additionally, we offer a compilation of recommendations for future research that emerged during our investigation.

6.1 Conclusion

The purpose of this study was firstly to understand the drug development industry's perspective on ethical dilemmas which arise from using black box AI models. Secondly, we wanted to compare this perspective to current EU requirements related to AI use, specifically on the areas related to black box AI. Additionally, this study aims to understand the industry's perspective on ethical dilemmas arising from using black box AI models and to compare these perspectives with current EU requirements specifically related to black box AI.

The drug development industry champions AI with fervor, driven by its potential to revolutionize processes. There were varying perspectives on how much AI innovation needs to consider ethical aspects, however, there's consensus: patient well-being must never be compromised for technological advancement. To navigate this complex area, the industry advocates for direction and clarity in form of requirements and regulations, particularly regarding AI explainability.

EU requirements used as frameworks in this study prioritize ethics possibly at the expense of technological progress, with stringent requirements that could be perceived as limiting the exploration of AI's full potential. These requirements align with the industry perspective when it comes to balancing innovation and ethics. Moreover, EU requirements underscore the critical role of human participation in deploying AI in medical devices, recognizing such applications as high-risk. This requirement of human involvement from the EU aligns with the perspective of the drug development industry.

The alignment suggests that both the requirement frameworks and the industry are moving towards a common goal of safe, ethical, and effective integration of AI technologies in drug development. We see a correspondence between the EU frameworks and the drug development industry's perspective. We can also ascertain that the drug development industry expresses need and shows acceptance of requirements regarding the use of black box AI, and AI use in general.

6.2 Limitations

The study was conducted on a rather small selection of participants, which is a product of the challenge of getting in contact with participants within the rather specific field, who were willing and comfortable with participating in an interview. We also created the questions for the interviews with terminology which we assessed to be relevant based on our literature review, such as the term ‘black box’. However, we noticed that the participants in the interviews were in several cases not familiar with the terminology. This was a consequence of our unawareness of the common language within the industry, and could have been avoided if we had discussed terminology with a representative from the drug development industry prior to conducting the interviews.

6.3 Recommendations for Future Research

The findings of this study sheds light on areas where further research could advance our understanding of the ethical implications of AI in the drug development industry. Future studies should consider a broader selection of industries and geographic locations to determine whether the ethical concerns and trust issues identified are universally relevant or vary significantly across different contexts. This could involve comparative studies that examine how different regulatory environments impact the adoption and perception of AI technologies

To address the emergent trust issues between AI users and stakeholders revealed in our interviews, particularly concerning the explainability of AI systems, future research could conduct in-depth studies exploring how transparency and ethical considerations influence trust dynamics. This research is essential to creating policies and procedures that improve the ethical governance, predictability, and dependability of AI technology in the market.

References

Adeoye-Olatunde, O. A. & Olenik, N. L. (2021). Research and Scholarly Methods: Semi-Structured Interviews, *JACCP: JOURNAL OF THE AMERICAN COLLEGE OF CLINICAL PHARMACY*, [e-journal] vol. 4, no. 10, pp.1358–1367, Available Online: https://onlinelibrary.wiley.com/doi/abs/10.1002/jac5.1441 [Accessed 17 April 2024]

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P. & Zhavoronkov, A. (2016). Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data, *Molecular Pharmaceutics*, [e-journal] vol. 13, no. 7, pp.2524–2530, Available Online: https://pubs.acs.org/doi/10.1021/acs.molpharmaceut.6b00248 [Accessed 24 April 2024]

Anjomshoae, S. (2022). Context-Based Explanations for Machine Learning Predictions, Umeå: Department of Computing Science, Umeå University

Arreche, O., Guntur, T. R., Roberts, J. W. & Abdallah, M. (2024). E-XAI: Evaluating Black-Box Explainable AI Frameworks for Network Intrusion Detection, *IEEE Access*, [e-journal] vol. 12, pp.23954–23988, Available Online: https://ieeexplore.ieee.org/document/10433134/ [Accessed 25 March 2024]

Besold, T. R. & Uckelman, S. L. (2018). The What, the Why, and the How of Artificial Explanations in Automated Decision-Making, [e-journal], Available Online: https://arxiv.org/abs/1808.07074 [Accessed 9 May 2024]

BioStock (2023) Drug development: The four phases. [Online] Available online: https://www.biostock.se/en/2023/01/drug-development-the-four-phases/ [Accessed 15 April 2024]

Blanco-González, A., Cabezón, A., Seco-González, A., Conde-Torres, D., Antelo-Riveiro, P., Piñeiro, Á. & Garcia-Fandino, R. (2023). The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies, *Pharmaceuticals*, [e-journal] vol. 16, no. 6, p.891, Available Online: https://www.mdpi.com/1424-8247/16/6/891 [Accessed 13 April 2024]

Braun, V. & Clarke, V. (2006). Using Thematic Analysis in Psychology, *Qualitative Research in Psychology*, [e-journal] vol. 3, no. 2, pp.77–101, Available Online: http://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa [Accessed 24 April 2024]

Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms, *Big Data & Society*, [e-journal] vol. 3, no. 1, p.205395171562251, Available Online: http://journals.sagepub.com/doi/10.1177/2053951715622512 [Accessed 24 April 2024]

Cassauwers, T. (2020). Opening the 'Black Box' of Artificial Intelligence | Research and Innovation, Available Online: https://projects.research-and-innovation.ec.europa.eu/en/horizon-magazine/opening-black-box-artificial-intelligence [Accessed 21 April 2024]

Chen, Z., Liu, X., Hogan, W., Shenkman, E. & Bian, J. (2021). Applications of Artificial Intelligence in Drug Development Using Real-World Data, *Drug Discovery Today*, [e-journal] vol. 26, no. 5, pp.1256–1264, Available Online: https://linkinghub.elsevier.com/retrieve/pii/S1359644620305316 [Accessed 20 April 2024]

Chui, M., Hall, B., Mayhew, H., Singla, A., Sukharevsky, A., Angalakudati, S., Bankhwal, M., DeLallo, D., Hanselman, H., Patel, V., & Wang, W. (2022). The State of AI in 2022 - and a half decade in review. Available Online: https://www.mckinsey.com/~/media/mckinsey/business%20functions/quantumblack/our%20insights/the%20state%20of%20ai%20in%202022%20and%20a%20half%20decade%20in%20review/the-state-of-ai-in-2022-and-a-half-decade-in-review.pdf [Accessed 10 April 2024]

Daly, J., Kellehear, A. & Gliksman, M. D. (1997). The Public Health Researcher: A Methodological Guide, New York: Oxford university press. [Accessed 24 April 2024]

Davids, J., Lidströmer, N. & Ashrafian, H. (2022). Artificial Intelligence in Medicine Using Quantum Computing in the Future of Healthcare, in N. Lidströmer & H. Ashrafian (eds), *Artificial Intelligence in Medicine*, [e-book] Cham: Springer International Publishing, pp.423–446, Available Online: https://doi.org/10.1007/978-3-030-64573-1_338 [Accessed 23 April 2024]

Durán, J. M. & Jongsma, K. R. (2021). Who Is Afraid of Black Box Algorithms? On the Epistemological and Ethical Basis of Trust in Medical AI, *Journal of Medical Ethics*, p.medethics-2020-106820

Emerson, R. W. (2015). Convenience Sampling, Random Sampling, and Snowball Sampling: How Does Sampling Affect the Validity of Research?, *Journal of Visual Impairment & Blindness*, [e-journal] vol. 109, no. 2, pp.164–168, Available Online: http://journals.sagepub.com/doi/10.1177/0145482X1510900215 [Accessed 11 April 2024]

European Commission. Directorate General for Communications Networks, Content and Technology. (2020). The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment., [e-book] LU: Publications Office, Available Online: https://data.europa.eu/doi/10.2759/791819 [Accessed 9 April 2024]

European Commission. (2024). AI Act | Shaping Europe's Digital Future, Available Online: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai [Accessed 3 May 2024]

European Data Protection Supervisor. (2023). TechDispatch: Explainable Artificial Intelligence., [e-book] LU: Publications Office, Available Online: https://data.europa.eu/doi/10.2804/802043 [Accessed 10 April 2024]

European Medicines Agency. (2023). Draft reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. [pdf] Available Online: https://www.ema.europa.eu/en/documents/scientific-guideline/draft-reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle_en.pdf [Accessed 13 April 2024]

European Parliament. (2023). EU AI Act: First Regulation on Artificial Intelligence. *Topics | European Parliament*, Available Online: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [Accessed 8 May 2024]

European Parliament and Council of the European Union. (2024) Regulation (EU) 2024/... on harmonised rules on artificial intelligence and amending Regulations and Directives (Artificial Intelligence Act), P9_TA(2024)0138, Official Journal of the European Union, Available online: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf [Accessed 1 May 2024]

Franklin, C. & Ballan, M. (2001). Reliability and validity in qualitative research. In: B. Thyer, ed., The Handbook of Social Work Research Methods, 1st ed. Thousand Oaks, CA: Sage Publications, pp. 273-292.

Furlong, A. (2023). Regulating the Machine: Europe's Race to Get to Grips with AI Drugs, *POLITICO*, Available Online: https://www.politico.eu/article/regulate-europe-race-artificial-intelligence-ai-drugs-medicines/ [Accessed 27 April 2024]

Gbadegeshin, S. A., Al Natsheh, A., Ghafel, K., Tikkanen, J., Gray, A., Rimpiläinen, A., Kuoppala, A., Kalermo-Poranen, J. & Hirvonen, N. (2021). WHAT IS AN ARTIFICIAL INTELLIGENCE (AI): A SIMPLE BUZZWORD OR A WORTHWHILE INEVITABILITY?, 14th Annual International Conference of Education, Research and Innovation, Online Conference, November 2021, Online Conference, pp.468–479, Available Online: https://library.iated.org/view/GBADEGESHIN2021WHA [Accessed 9 April 2024]

Golafshani, N. (2015). Understanding Reliability and Validity in Qualitative Research, *The Qualitative Report*, [e-journal], Available Online: https://nsuworks.nova.edu/tqr/vol8/iss4/6/ [Accessed 22 April 2024]

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. & Yang, G.-Z. (2019). XAI—Explainable Artificial Intelligence, *Science Robotics*, [e-journal] vol. 4, no. 37, p.eaay7120, Available Online: https://www.science.org/doi/10.1126/scirobotics.aay7120 [Accessed 24 April 2024]

Gyevnar, B., Ferguson, N. & Schafer, B. (2023). Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act?, [e-book], Available Online: http://arxiv.org/abs/2302.10766 [Accessed 6 May 2024]

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M. & Hussain, A. (2024). Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence, *Cognitive Computation*, [e-journal] vol. 16, no. 1, pp.45–74, Available Online: https://link.springer.com/10.1007/s12559-023-10179-8 [Accessed 24 April 2024]

Humphreys, P. (2009). The Philosophical Novelty of Computer Simulation Methods, *Synthese*, vol. 169, no. 3, pp.615–626

Iniesta, R. (2023). What Can Humans Do to Guarantee an Ethical AI in Healthcare? Part II | Feature from King's College London, Available Online: https://www.kcl.ac.uk/what-can-humans-do-to-guarantee-an-ethical-ai-in-healthcare-part-ii [Accessed 27 April 2024]

Joppe, M. (2000). The research process. Available at: http://www.ryerson.ca/~mjoppe/rp.htm [Accessed 19 April 2024].

Kalodanis, K., Rizomiliotis, P. & Anagnostopoulos, D. (2023). European Artificial Intelligence Act: An AI Security Approach, *Information & Computer Security*, [e-journal], Available Online: https://www.emerald.com/insight/content/doi/10.1108/ICS-10-2022-0165/full/html [Accessed 6 May 2024]

Kriukow, J. (2024) The Scholar's Guide to AI-assisted Thematic Analysis [PDF] Available Online: https://payhip.com/b/kmzOL [Accessed: 6 May 2024]

Wang, V. X. (ed.). (2015). Handbook of Research on Scholarly Publishing and Research Methods:, [e-book] IGI Global, Available Online: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-4666-7409-7 [Accessed 12 May 2024]

Lepri, B., Oliver, N., Letouzé, E., Pentland, A. & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges, *Philosophy & Technology*, [e-journal] vol. 31, no. 4, pp.611–627, Available Online: http://link.springer.com/10.1007/s13347-017-0279-x [Accessed 24 April 2024]

Leslie, D. (2019). Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector, [object Object], Available Online: https://zenodo.org/record/3240529 [Accessed 17 April 2024]

Läkemedelsverket. (2023). Användning av artificiell intelligens i svensk sjukvård | Läkemedelsverket, Available Online: https://www.lakemedelsverket.se/sv/medicinteknik/anvanda/anvandning-av-ai-i-sjukvarden [Accessed 9 April 2024]

McDermid, J. A., Jia, Y., Porter, Z. & Habli, I. (2021). Artificial Intelligence Explainability: The Technical and Ethical Dimensions, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, [e-journal] vol. 379, no. 2207, p.20200363, Available Online: https://royalsocietypublishing.org/doi/10.1098/rsta.2020.0363 [Accessed 6 May 2024]

MedTech Europe. (2024). *Medical technology industry perspective on the final AI Act*. Available Online: https://www.medtecheurope.org/resource-library/medical-technology-industry-perspective-final-ai-act/ [Accessed 7 May 2024]

Miller, T., Howe, P. & Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences, [e-journal], Available Online: https://arxiv.org/abs/1712.00547 [Accessed 10 April 2024]

Nannini, L., Balayn, A. & Smith, A. L. (2023). Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK, in *2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’23: The 2023 ACM Conference on Fairness, Accountability, and Transparency, Chicago IL USA, 12 June 2023, Chicago IL USA: ACM, pp.1198–1212, Available Online: https://dl.acm.org/doi/10.1145/3593013.3594074 [Accessed 6 May 2024]

Oates, B. J. (2006). Researching Information Systems and Computing, 1. publ., Repr., Los Angeles: SAGE

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J. & Gomez, E. (2023). The Role of Explainable AI in the Context of the AI Act, in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 12 June 2023, New York, NY, USA: Association for Computing Machinery, pp.1139–1150, Available Online: https://dl.acm.org/doi/10.1145/3593013.3594069 [Accessed 6 May 2024]

Patheon (2023) Drug development phases. [Online] Available Online t: https://www.patheon.com/us/en/insights-resources/blog/drug-development-phases.html [Accessed 15 April 2024]

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K. & Tekade, R. K. (2021). Artificial Intelligence in Drug Discovery and Development, *Drug Discovery Today*, [e-journal] vol. 26, no. 1, pp.80–93, Available Online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7577280/ [Accessed 23 April 2024]

Peters, U. (2023). Explainable AI Lacks Regulative Reasons: Why AI and Human Decision-Making Are Not Equally Opaque, *AI and Ethics*, [e-journal] vol. 3, no. 3, pp.963–974, Available Online: https://link.springer.com/10.1007/s43681-022-00217-w [Accessed 24 April 2024]

Petersson, L., Larsson, I., Nygren, J. M., Nilsen, P., Neher, M., Reed, J. E., Tyskbo, D. & Svedberg, P. (2022). Challenges to Implementing Artificial Intelligence in Healthcare: A Qualitative Interview Study with Healthcare Leaders in Sweden, *BMC Health Services Research*, [e-journal] vol. 22, no. 1, p.850, Available Online: https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-022-08215-8 [Accessed 18 April 2024]

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A. & Przybocki, M. A. (2021). Four Principles of Explainable Artificial Intelligence, NIST IR 8312, Gaithersburg, MD: National Institute of Standards and Technology (U.S.), p.NIST IR 8312, Available Online: https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf [Accessed 24 April 2024]

Roy, A., R. McDonald, P., Sittampalam, S. & Chaguturu, R. (2010). Open Access High Throughput Drug Discovery in the Public Domain: A Mount Everest in the Making, *Current Pharmaceutical Biotechnology*, [e-journal] vol. 11, no. 7, pp.764–778, Available Online: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1389-2010&volume=11&issue=7&spage=764 [Accessed 18 April 2024]

Ruedig, C. & Kler, G. (2022). Is the Time Ripe for AI Drug Discovery?, *AlbionVC*, Available Online: https://albion.vc/spotlight/deep-dives/is-the-time-ripe-for-ai-drug-discovery/ [Accessed 7 May 2024]

Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. (2014). Computational Methods in Drug Discovery, *Pharmacological Reviews*, [e-journal] vol. 66, no. 1, pp.334–395, Available Online: http://pharmrev.aspetjournals.org/lookup/doi/10.1124/pr.112.007336 [Accessed 18 April 2024]

Sovrano, F., Sapienza, S., Palmirani, M. & Vitali, F. (2022). Metrics, Explainability and the European AI Act Proposal, *J*, [e-journal] vol. 5, no. 1, pp.126–138, Available Online: https://www.mdpi.com/2571-8800/5/1/10 [Accessed 6 May 2024]

Stahl, B. C., Antoniou, J., Ryan, M., Macnish, K. & Jiya, T. (2022). Organisational Responses to the Ethical Issues of Artificial Intelligence, *AI & SOCIETY*, [e-journal] vol. 37, no. 1, pp.23–37, Available Online: https://link.springer.com/10.1007/s00146-021-01148-6 [Accessed 25 March 2024]

Stenbacka, C. (2001). Qualitative Research Requires Quality Concepts of Its Own, *Management Decision*, [e-journal] vol. 39, no. 7, pp.551–556, Available Online: https://www.emerald.com/insight/content/doi/10.1108/EUM0000000005801/full/html [Accessed 23 April 2024]

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R. & Collins, J. J. (2020). A Deep Learning Approach to Antibiotic Discovery, *Cell*, vol. 180, no. 4, pp.688-702.e13 [Accessed 9 April 2024]

SwedenBIO. (2023). The Swedish Drug Discovery and Development Pipeline Report 2023 , Available Online: https://www.mediconvillage.se/app/uploads/2023/03/Report_Swedish-Drug-Discovery-Development-Pipeline_2023.pdf [Accessed 11 April 2024]

Tamimi, N. A. M. & Ellis, P. (2009). Drug Development: From Concept to Marketing!, *Nephron Clinical Practice*, [e-journal] vol. 113, no. 3, pp.c125–c131, Available Online: https://www.karger.com/Article/FullText/232592 [Accessed 12 April 2024]

Thampi, A. (2022). Interpretable AI: Building Explainable Machine Learning Systems, Shelter Island, NY: Manning Publications

Tonkens, R. (2005). An Overview of the Drug Development Process, *Physician Executive*, vol. 31, no. 3, pp.48–52. Available Online: link.gale.com/apps/doc/A133219960/AONE?u=googlescholar&sid=bookmark-AONE&xid=88d2a55a [Accessed 10 April 2024]

Tripathi, M. (2021). Sentiment Analysis of Nepali COVID19 Tweets Using NB, SVM AND LSTM, *Journal of Artificial Intelligence and Capsule Networks*, vol. 3, no. 3, pp.151–168 [Accessed 12 April 2024]

U.S. Food and Drug Administration (2018) Drug development process. [Online] Available Online:https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process [Accessed 15 April 2024]

Wainer, H. & Braun, H. I. (1988). Test Validity, Hillsdale, N.J.: L. Erlbaum Associates

Weidener, L. & Fischer, M. (2024). Role of Ethics in Developing AI-Based Applications in Medicine: Insights From Expert Interviews and Discussion of Implications, *JMIR AI*, [e-journal] vol. 3, p.e51204, Available Online: https://ai.jmir.org/2024/1/e51204 [Accessed 17 April 2024]

World Health Organization (WHO) (n.d.). Available Online: https://www.who.int/ [Accessed 23 April 2024]

World Health Organization. (n.d.). Guiding optimal development and use of digital health towards improved health outcomes, Available Online: https://www.who.int/westernpacific/activities/guiding-optimal-development-and-use-of-digital-health-towards-improved-health-outcomes [Accessed 9 June 2024]

Appendix A - Interview Guide

Phase 1: Introduction (with focus on the research) and ethical review

We are studying the attitude and perspective of the drug development industry regarding ethical concerns with unexplainable AI. Perhaps you are familiar with the Assessment List for Trustworthy AI that has been developed by the EU? In the list they give some guidelines for the use of AI, not specifically for healthcare but for all organizations utilizing AI systems. In this list they discuss transparency and accountability quite a lot. What we are researching is how well drug development companies' perspectives on AI align with the EU ALTAI list when it comes to explainability of AI models.

We are not studying how well the work of the companies align with the requirements, rather how well the contents of the ALTAI list, when it comes to explainability/transparency, aligns with the reality of the drug development industry.

We will be referring to unexplainable AI as “Black Box” AI, and we will be referring to AI which is not “Black Box” as explainable AI (XAI).

Phase 2: Background (with focus on the respondent)

1. Can you please introduce yourself and what you work with/research?

   a. Can you describe your relation to drug development?

   b. How does AI intersect with your work/research?

Phase 3: Exploring Ethical Challenges

2. Is the discussion of Explainable Artificial Intelligence (XAI) something that is common within the field of your work?

   a. Have you encountered or utilized "black box" AI models in your work?

3. What ethical implications do you see with AI in drug development?

   a. What do you regard as the most problematic ethical issue with Black Box AI models?

   b. Do you regard explainability as an ethical issue with AI?

   c. Do you regard traceability as an ethical issue with AI?

Phase 4: Addressing Ethical Implications

4. Do the ethical challenges with AI affect your work/your company's work?

   a. How do you personally believe that these ethical challenges can be addressed?

5. In your opinion, is it essential to prioritize explainability and traceability in AI models specifically within the context of drug development?

Phase 5: Exploring Opportunities and recommendations

6. In your opinion, what are the biggest potential benefits or opportunities that could arise from effectively addressing the ethical challenges of AI in drug development?

   a. Would they be possible without black box AI in any way?

Phase 6: Wrap-up and finalization

7. Do you have anything else that you would like to add regarding your own perspective on AI that you think would be valuable for our research?

Appendix B - Informed Consent Form

Information to interviewees

We would like to ask you if you would be willing to participate in a research project. In this document, we will give you information about the project and what it means to participate and be a part of this research.

What kind of project is this and why do we reach out to you as a participant

This project investigates the drug development industry's perspective on the ethical concerns posed by AI and AI black box. The research also focuses on the EU requirements and the drug development industry's attitude and position towards those requirements. Because of the fact that you work within the medicine and/or drug development industry, you have been chosen as a possible interviewee.

The research is conducted by bachelor students at the institution of informatics at Lunds University.

How is the project performed

The study is performed as 4-6 semi-structured qualitative interviews with individuals currently working in the drug development/medicine industry. The length of the interviews are approximately 20-40 minutes and can be performed digitally or face to face. You will find the interview questions attached down below in this information document

We will handle your interview data in a way that protects your privacy. Interview data is only fully accessible to the project's principal investigator.

You have the right to participate in the project using your own data and to remove it from it. Should you desire to do so, please get in touch with the project's lead researcher.

How will you obtain information about the interview data and the result of the project

The interviewees can obtain a summary of findings on request.

Participation is voluntary

It is optional to participate.You are free to withdraw from the project at any time without having to give a reason. Your involvement in it is entirely voluntary. Should you decide to withdraw from the project, kindly get in touch with one of the project's relevant researchers.

Responsible researchers

Sofie Schnitzer, sofie.schnitzer@gmail.com, bachelor student, department of informatics, Lund University School of Economics and Management, Sweden

Mathewos Tiruwuha, mathewostiruwuha@gmail.com, bachelor student, Department of informatics, Lund University School of Economics and Management, Sweden

Consent to participate in the research project

I have obtained written information about the study and have had an opportunity to ask questions. I will keep the written information.

- I give my consent to participate as an interviewee in the research project *'Transparency in the Black box; The drug development industry's perspective on ethical concerns posed by AI, and how well it aligns with EU requirements'.*

____

Place and date

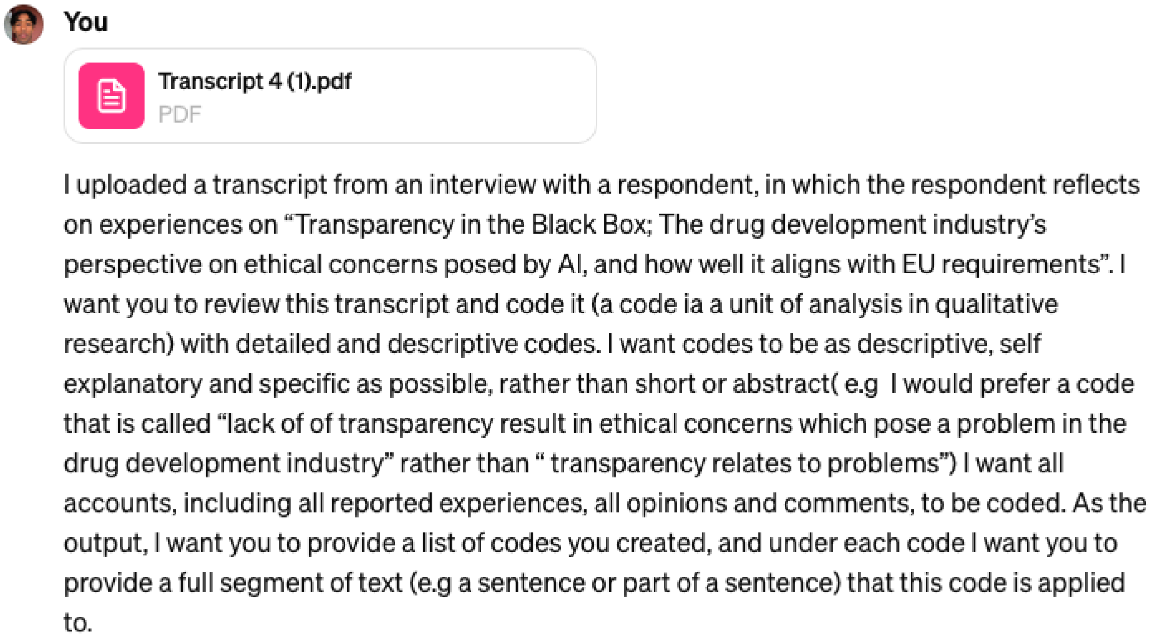
____

Signature

____

Name clarification

Appendix H - ChatGPT Prompt 1

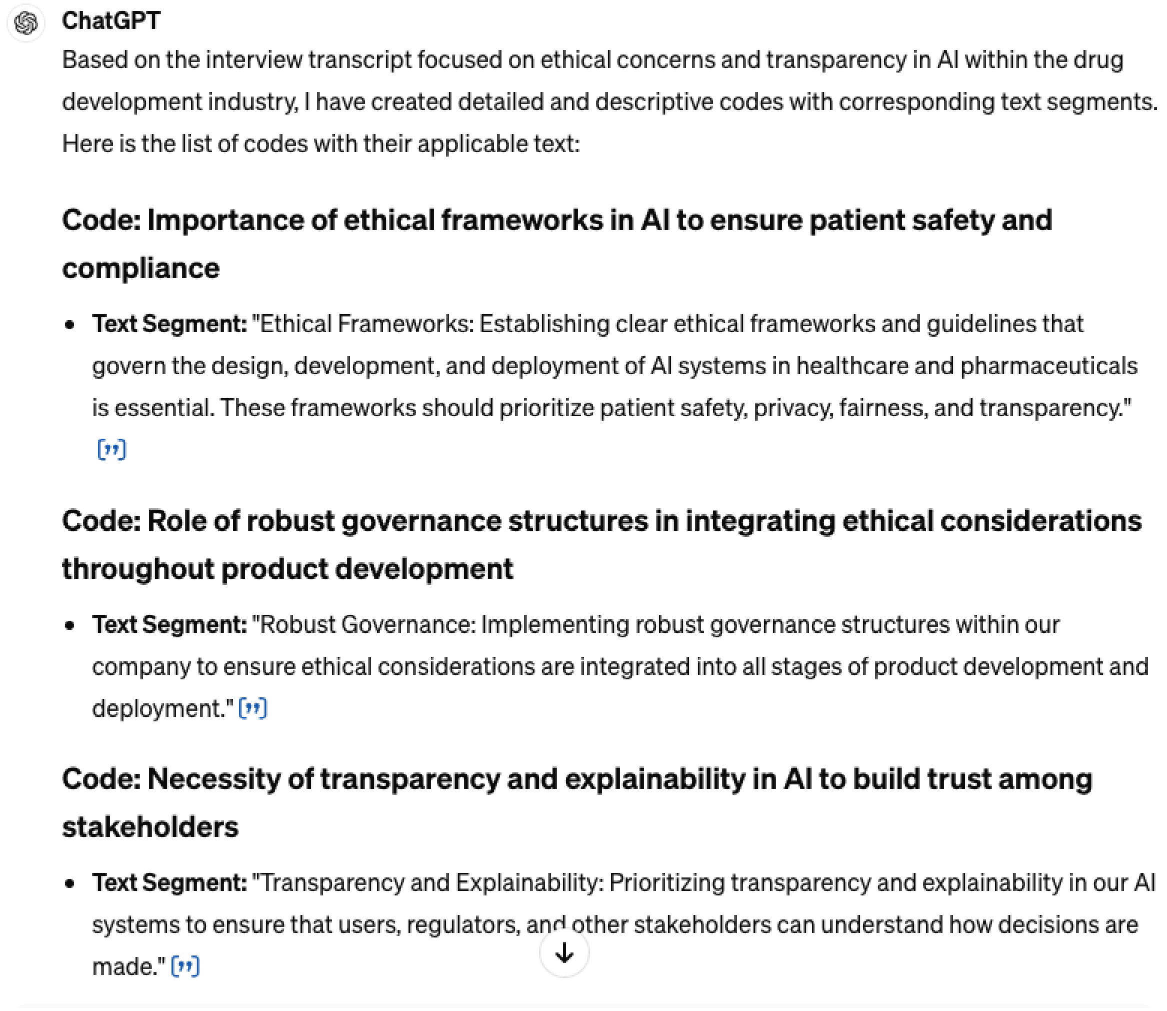


You

Transcript 4 (1).pdf
PDF

I uploaded a transcript from an interview with a respondent, in which the respondent reflects on experiences on “Transparency in the Black Box; The drug development industry’s perspective on ethical concerns posed by AI, and how well it aligns with EU requirements”. I want you to review this transcript and code it (a code ia a unit of analysis in qualitative research) with detailed and descriptive codes. I want codes to be as descriptive, self explanatory and specific as possible, rather than short or abstract(e.g I would prefer a code that is called “lack of of transparency result in ethical concerns which pose a problem in the drug development industry” rather than “ transparency relates to problems”) I want all accounts, including all reported experiences, all opinions and comments, to be coded. As the output, I want you to provide a list of codes you created, and under each code I want you to provide a full segment of text (e.g a sentence or part of a sentence) that this code is applied to.

Appendix I - ChatGPT Output



ChatGPT

Based on the interview transcript focused on ethical concerns and transparency in AI within the drug development industry, I have created detailed and descriptive codes with corresponding text segments. Here is the list of codes with their applicable text:

Code: Importance of ethical frameworks in AI to ensure patient safety and compliance

- **Text Segment:** "Ethical Frameworks: Establishing clear ethical frameworks and guidelines that govern the design, development, and deployment of AI systems in healthcare and pharmaceuticals is essential. These frameworks should prioritize patient safety, privacy, fairness, and transparency."

Code: Role of robust governance structures in integrating ethical considerations throughout product development

- **Text Segment:** "Robust Governance: Implementing robust governance structures within our company to ensure ethical considerations are integrated into all stages of product development and deployment."

Code: Necessity of transparency and explainability in AI to build trust among stakeholders

- **Text Segment:** "Transparency and Explainability: Prioritizing transparency and explainability in our AI systems to ensure that users, regulators, and other stakeholders can understand how decisions are made."

Appendix J - AI Contribution Statement

In the completion of our bachelor thesis, titled "Transparency in the Black Box: The drug development industry's perspective on ethical concerns posed by AI, and how well it aligns with EU requirements," we employed two principal AI-based tools: ChatGPT and Whisper.

Tools Utilized:

1. *ChatGPT:* Utilized extensively throughout the thesis for thematic analysis, grammatical improvements, and sentence structuring. ChatGPT assisted in refining the language and clarity of our arguments across all sections of the document.

2. *Whisper:* Employed solely for transcription purposes. Whisper was used to convert audio recordings of interviews into text, which were then analyzed within the thesis.

Degree of Utilization:

- *ChatGPT:* The use of ChatGPT was focused primarily on refining the language of the entire thesis. It provided support in grammar corrections and sentence structuring across all sections, enhancing the readability and academic tone of the document. ChatGPT's role was crucial in the thematic analysis phase, where it helped in coding and color-coding the data, ensuring that thematic consistency was maintained throughout the analysis.

- *Whisper:* Its functionality was confined to the transcription of recorded interviews, which formed the empirical data base for our analysis. This accurate transcription was critical for the integrity of our qualitative analysis.

Our research procedures were made more efficient and our study's analytical depth was improved by the incorporation of these AI techniques. Specifically, ChatGPT played a key role in refining the thesis's wording, which improved the way our findings were presented and made sense. Conversely, Whisper made sure that the data analysis we conducted was supported by precisely recorded empirical evidence.

We have openly acknowledged these AI tools' contributions in this thesis, showing how they have improved and supported our research process and results.