.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Cruz, Maria Renata Keithlyn de Gois.
Leaflet for Artificial Intelligence Systems (LAIS): a novel artefact for promoting transparency in AI / Maria Renata Keithlyn de Gois Cruz. - Recife, 2025.
96f.: il.

Dissertação (Mestrado) - Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciência da Computação, 2025.
Orientação: Carina Alves Frota.
Coorientação: Geber Ramalho.
Inclui referências e apêndices.

1. Artificial Intelligence; 2. Ethical AI; 3. Leaflet; 4. Responsible AI. I. Alves, Carina Frota. II. Ramalho, Geber. III. Título.

UFPE-Biblioteca Central

UNIVERSIDADE FEDERAL DE PERNAMBUCO CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MARIA RENATA KEITHLYN DE GOIS CRUZ

Leaflet for Artificial Intelligence Systems (LAIS): A Novel Artefact for Promoting Transparency in AI

RECIFE, 2025

MARIA RENATA KEITHLYN DE GOIS CRUZ

Leaflet for Artificial Intelligence Systems (LAIS): A Novel Artefact for Promoting Transparency in AI

Esta dissertação foi apresentada à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

ORIENTADORA: PROF. DRA. CARINA FROTA ALVES
CO-ORIENTADOR: PROF. DR. GEBER RAMALHO

RECIFE, 2025

Every New Tool of Technology Creates New Types of Failures- Shannon French

ACKNOWLEDGEMENTS

First, I would like to thank **God** for giving me this incredible opportunity. HE only gives what we can bear, therefore, I am grateful for receiving such abilities. To Our Holy Lady Mary, who covered me with Her Sacred Clock, and interceded for me. *Saint Benedict, Saint Joseph, Saint Gianna*, who encouraged me throughout this journey.

Secondly, I would like to thank my parents, *Andréa and Rogério*, who have always supported me, even though the path I have chosen is not easy nor amicable. Thirdly, I would like to thank my husband, *Victor Taka*, who gave and continues to give me, courage and support in horrendous times. I would not be able to do this without your support.

Also, I am grateful to my advisors: *Professor Carina*, thank you for slowing me down and shaping me with your iron first. I hope I have achieved your standards.

And to Professor *Geber Ramalho*, you are a treasure that I have found in this life. Thank you for giving me hope when I was in the desert and for being optimistic even in my biggest failures. I must say you are an extraordinary and visionary man; I look up to you, and I am tremendously grateful to be your advisee.

To all my friends, colleagues, and research pals, especially **Sophia Calado and Diego Madeira**. The path is not easy but, without you: It would have been impossible.

> **Not many of you should become masters, my fellow believers, because you know that we who teach will be judged more strictly. (James, 3; 1)**

ABSTRACT

The adoption of AI (Artificial Intelligence) systems by the general population led to many juridical and ethical problems. Researchers have proposed a set of widely accepted ethical principles to guide the development of AI systems. However, it is not clear how to translate the principles into practice. In this dissertation we aim to address the transparency principle and investigate how the lack of this principle may impact on the operation of AI systems, leading to problems such as the unclearness of who is responsible when a negative output is generated. Moreover, we aim to explore the viewpoints of developers as well as users of AI systems, as they often do not understand nor know the system's limitations and implications. We propose a *Leaflet for Artificial Intelligence Systems* (LAIS), an artefact that aims to support transparency in AI systems. This artefact was crafted inspired by the pharmaceutical leaflets, and it was developed due to the recognition that the current software licenses are lengthy and not intuitive. We adopted the Design Science Research Methodology to guide the development of the artefact. To design the leaflet, we analyzed AI regulations initiatives, ethical AI literature, and pharmaceutical leaflets. To validate our artefact, we conducted a survey with experts from different backgrounds, including developers, users and researchers in the field of AI. The LAIS achieved a 94% approval rating, with participants affirming its relevance as a document that fosters transparency. Therefore, we consider that the LAIS represents an original and innovative contribution which operationalizes transparency principle into a concrete artefact, and it was crafted inspired by a long tradition of the pharmaceutical industry in communicating with end-users.

Key-Words: Artificial Intelligence, Transparency, Ethical AI, Leaflet.

RESUMO

A adoção de sistemas de IA (Inteligência Artificial) pela população em geral tem gerado muitos problemas jurídicos e éticos. Para orientar o desenvolvimento de sistemas de IA, pesquisadores propuseram um conjunto de princípios éticos amplamente aceitos. No entanto, não está claro como traduzir esses princípios na prática. Nesta dissertação, busca-se abordar o princípio ético da transparência e investigar como a ausência desse princípio pode impactar na operação dos sistemas de IA ocasionando problemas como a indefinição do responsável quando um resultado negativo é gerado pelo sistema. Além disso, pretende-se explorar os pontos de vista de desenvolvedores e usuários de sistemas de IA, já que frequentemente os usuários não compreendem nem conhecem as limitações e implicações desses sistemas. Como contribuição, foi proposta a Bula para Sistemas de Inteligência Artificial (LAIS, na sigla em inglês), um artefato que tem como objetivo apoiar a transparência em sistemas de IA. Este artefato foi inspirado nas bulas farmacêuticas e foi desenvolvido diante do reconhecimento que as atuais licenças de software são extensas e pouco intuitivas. Adotou-se a Metodologia de Pesquisa Design Science Research para fundamentar o desenvolvimento do artefato. Para projetar a Bula, analisou-se iniciativas de regulamentação de IA, literatura sobre ética em IA e bulas farmacêuticas. Para validar o artefato, foi realizado um survey com especialistas de diferentes áreas, incluindo desenvolvedores, usuários e pesquisadores no campo da IA. A LAIS alcançou uma aprovação de 94%, com os participantes afirmando que ela é relevante para a promoção da transparência nos sistemas de IA. Portanto, considera-se que a LAIS é uma contribuição original e inovadora, que operacionaliza o princípio da transparência em um artefato concreto, inspirado no design de bulas farmacêuticas e que facilita na comunicação e compreensão dos usuários finais.

Palavras-Chave: Inteligência Artificial, Transparência, Ética em IA, Bula Farmacêutica

SUMMARY

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

AI- Artificial Intelligence Systems

ANVISA - Agência Nacional de Vigilância Sanitária

DSR- Design Science Research

DSRM-Design Science Research Methodology

FDA- Food and Drug Administration

GDPR- General Data Protection Regulation

GPL- GNU General Public License

LAIS- Leaflet for Artificial Intelligence Systems

LGPD- Lei Geral de Proteção de Dados

MIT- Massachusetts Institute of Technology

NHTSA- National Highway Traffic Safety Administration

RAI- Responsible Artificial Intelligence Systems

RAIL- Responsible AI Licenses

TGA- Therapeutic Goods Administration

XAI- Explainable Artificial Intelligence Systems

1. INTRODUCTION

In recent years, the wide adoption of AI systems by the general population has given rise to several juridical and ethical issues. The utilization of these systems in daily life have amplified the negative outputs generated by AI Systems, as the users often do not understand nor know the system's limitations and implications. An example of these conundrums is autonomous car crashes, such as Tesla's 2021 case.

In 2021, Tesla's Autopilot system was involved in a significant incident that resulted in 13 severe injuries, raising concerns about the responsibility and safety of autonomous driving technology. Investigations by the National Highway Traffic Safety Administration (NHTSA) reviewed 956 crashes linked to Tesla vehicles from August 2018 to August 2023. The agency concluded that in approximately half of these cases, there was either insufficient data to assess the situation, or Tesla’s Autopilot was not engaged at the time of the crash, leaving the other vehicle at fault or the incident unrelated to the probe. In 211 incidents, Tesla's Autopilot failed to prevent a frontal collision with another vehicle or obstacle, even when an attentive driver could have mitigated the crash. Additionally, 145 crashes occurred under low-traction conditions, such as wet roads, while 111 incidents were attributed to the unintentional disengagement of Autosteer[1] due to driver input. These accidents are often severe, as both the system and the driver fail to respond adequately, leading to high-speed crashes with substantial energy impact.

Despite Tesla's instructions for drivers to maintain attention and keep their hands on the wheel while using Autopilot, in this research we understand that the potential risks and possible negative consequences are not sufficiently transparent or clear for the users (Dignum, 2019). Therefore, we had an initial hypothesis that the AI industry could not be sufficiently transparent with the users, causing an increase in the number of negative outputs.

Most ethical AI guidelines converge on critical principles such as transparency, fairness/justice, responsibility, privacy, safety, and sustainability (Tieto, 2018) (Sony, 2018) (The Japanese Society for Artificial Intelligence Ethical Guidelines, 2017) (Microsoft, 2018)(The Public Voice, 2018). However, as Jedlickova (2024) states, the society urges not for ethical principles, but concrete artefacts: "*Ethical principles outlined in the guidelines should be further translated into concrete requirements (...)*

[1] Tesla’s Autopilot

[So] These requirements can be effectively reviewed during subsequent evaluation processes".

As dealing with all principles is a large scope, we investigate the transparency principle. This principle underpins critical attributes such as trust, explicability, and responsibility, as it may clarify the stakeholders' responsibilities when a negative output is generated. In the transparency scope, the current software licenses do not seem to be an adequate tool to establish an effective communication channel between developers and users. Indeed, software licenses are lengthy and not intuitive.

In this dissertation, we propose a *Leaflet for Artificial Intelligence Systems* (LAIS) to support transparency in AI systems. The artefact was inspired by pharmaceutical leaflets. There are other propositions of leaflets in the software domain, but they have different goals such as making more transparent the documentation of requirements (Lima, 2015) (Leal et al, 2012).

To design the leaflet, we conducted a thorough analysis of AI regulatory initiatives, the ethical AI literature, and a selection of pharmaceutical leaflets. To validate the AI leaflet, we administered a survey to experts from diverse fields. The Leaflet for Artificial Intelligence Systems (LAIS) achieved a 94% approval rating in the specialists' revision, reinforcing the original and innovative contribution of LAIS to translate the ethical transparency principle into a concrete artefact.

1.1 CONTEXT AND MOTIVATION

Our choice to focus on the transparency principle was due to the author's background in law. While we explored the current literature about Ethical AI, we noticed that although there are many studies about Ethical AI most of them are unclear about the stakeholders' responsibilities toward negative outputs (Deshpande et al.2022)

Nowadays, there is a convergence of five basic ethical principles for AI: transparency, fairness/justice, responsibility, privacy, safety, and sustainability (Jobin, 2019, Floridi, 2020). In our studies, we noticed a gap in the literature, the absence of applicable ethically aligned artefacts (Dignum, 2019).

Also, the presented principles try to support that AI systems operate in ways that align with ethical standards and societal values. While the transparency principle focuses on making the AI's processes, data use, and decision-making mechanisms clear and accessible to all stakeholders, fairness addresses the need to prevent bias and ensure equal treatment across diverse groups. Also, the privacy principle emphasizes safeguarding personal data, while accountability ensures that developers and operators can be held responsible for the AI's outcomes. While the safety principle ensures the system does not cause harm.

We noticed that the absence of transparency led to an increase in unresolved problems surrounding negative outputs generated by AI systems, such as what occurred in Tesla's case. Transparency, through explicability (Cortese, 2022), trust, and responsibility (Dignum, 2021) ensures that there is a clear assignment of responsibility when AI systems produce negative outcomes. It allows people affected by AI systems to identify, address, and rectify the causes of such events, as the users understand and know the limits of the system.

Then, we chose to focus specifically on the principle of transparency as it underpins critical elements such as trust, explicability, and affects directly in the definition of responsibility. Floridi (2020) argues that transparency is not merely about disclosing information but also ensuring that the rationale behind AI decisions can be explained intelligibly to non-expert audiences. Jobin (2019) considers explicability as a synonym of transparency. Notwithstanding, we understand that there are many sub-principles of transparency, as responsibility, accountability, sustantability, and rob. The latter a tool to promote transparency by making decisions understandable and fostering trust for the stakeholders involved (Jobin et al., 2019).

Also, the trust principle defended by Jobin (2019) is directly tied to transparency. She claims that for a system to be transparent, it needs the users to trust its operations, decision-making processes, and limitations. According to Jobin (2019), explicability is critical, because it offers clarity on the AI's functionality and allows users to question and verify outcomes. In contrast, trust refers to the confidence that stakeholders—such as developers, users, and regulators—place in the functionality, decision-making processes, and ethical integrity of AI systems. Provided that, Jobin (2019) claims that trust and explicability are core sub-principles to achieve transparency, as through them users understand how decisions are made. Also, the transparency principle relates closely to explicability and trust by ensuring users understand how decisions are made. At the same time, the responsibility principle focuses on clarifying who is accountable when AI systems produce negative outcomes.

1.2 OBJECTIVES

Software Licenses (e.g., the MIT License or Apache License 2.0) include disclaimers stating that the software is provided "as-is", however, it may not be effectively on their alert nor would guarantee its accuracy, as the warnings tend to be unclear or long. This example highlights a common problem in software licenses: they do not address the critical issues of responsibility in the event of negative outputs generated by AI systems. While these licenses describe the legal use of the software, they fail to provide ethical alignment, not holding someone responsible when adverse consequences occur. This gap underscores the need for new, transparent, and user-friendly artefacts. So, the *primary objective* of this research is to create an Leaflet aligned with the ethical principles of Artificial Intelligence (AI). We particularly aim to investigate the principle of transparency, as we aim to provide a clear, concise, and user-friendly artefact that outlines the key functionalities, risks, and stakeholder responsibilities of AI systems.

Inspired by the pharmaceutical industry, which has long dealt with similar issues of public trust, safety, and the clear definition of responsibilities, we developed the concept of a "Leaflet for Artificial Intelligence Systems" (LAIS). Similarly to the pharmaceutical leaflets designed to be transparent, the LAIS is a concise and user-friendly artefact that describes the risks, along with the stakeholder’s responsibilities.

The LAIS aims to provide clear information about the AI system's functionalities, risks, errors, and responsibilities. Also, this artefact aims to be a practical tool that helps the ethical alignment of AI systems while could mitigate the problems that arise from ambiguity in accountability and transparency (Dignum, 2020). The creation of the LAIS demonstrates how transparency, trust, and responsibility can be operationalized in AI systems, consequently, the LAIS ensures that users understand the AI system's operations and can hold responsible parties accountable when necessary.

The current dissertation’s contributions are a) a clear definition of stakeholder’s responsibilities when a negative output is generated by an AI system; b) an easy-to-understand, objective and user-friendly artefact ethically aligned with transparency principle for users and developers.

1.3 METHOD

In the initial stages of this research, the author’s academic background in law fostered a growing interest in the field of Ethical AI, prompting an exploration of the existing literature. This investigation revealed a significant gap in the allocation of responsibility for the outputs generated by AI systems. Numerous instances were identified where AI systems produced negative outcomes, yet no clear solutions or mechanisms for assigning responsibility were available to address these issues.

To further investigate, an ad hoc literature review was conducted to map the ethical principles governing AI systems, as can be seen in Chapter 2. Although numerous ethical principles were identified, their implementation and application in real-world contexts were inconsistent. This mapping process, combined with an analysis of the state of the art, highlighted two critical issues: the insufficient structure of the actual documents used regarding responsibilities in AI systems and the lack of clear transparency of these guidelines.

Drawing inspiration from the pharmaceutical industry's use of product leaflets, this research proposed the development of a similar artefact for AI systems, focusing on a transparent and understandable artefact. The resulting "Leaflet for Artificial Intelligence Systems" seeks to bridge key gaps in transparency while providing a user-friendly and accessible resource for stakeholders.

1.4 ROADMAP

This document is organized into eight chapters, including this introduction.

- Chapter 2 provides the definitions of the most relevant subjects for this research, such as Ethics, Ethical AI, Responsibility, and Transparency.
- Chapter 3 presents the research method used, its main phases and process, and the roadmap of the documents.
- Chapter 4 describes the Problem Characterization, explains why the lack of transparency is a problem, and what is expected from a solution.
- Chapter 5 discusses the related works, analysing the initiatives to improve transparency and accountability, and presents the literature review.
- Chapter 6 presents the proposed solution, its inspiration and explains how it was built.
- Chapter 7 discusses the validation of the artefact, through a survey conducted with researchers, practitioners and users of AI systems.

- Chapter 8 concludes this dissertation, discusses its limitations, threats to validity, contributions for the academy and industry, and proposes future works.

2. CONCEPTS

This Chapter defines the most relevant concepts to understand this dissertation. It overviews relevant subjects, such as Ethics, Ethical AI, Responsibility, and Transparency.

2.1 ETHICS

Ethics is the branch of philosophy that deals with questions of morality and principles of right and wrong behavior (ARISTÓTLES, 2011). Also, it encompasses the study of values, virtues, duties, and justice. It often seeks to establish guidelines for determining what actions are morally acceptable, how to balance competing moral claims, and how individuals and organizations should act in various contexts.

While often used interchangeably, ethics and morality refer to distinct yet related concepts, Ethics (Nalini, 2019) involves systematizing principles that guide behavior and help individuals or groups discern right or wrong. Conversely, morality refers to the inherent sense of right and wrong that individuals or societies possess, which often forms the foundation for ethical systems.

Utilitarianism is a traditional ethical theory that emphasizes the consequences of actions (Bentham, 1780; Miller, 2010). In this ethical theory, the utility of an artefact or action is often quantified through metrics such as satisfaction or dissatisfaction. Thus, utilitarianism often prioritizes the outcomes and consequences of actions, focusing on how these results contribute to the overall good or harm experienced by individuals or society.

In contrast, the deontological approach, which has its roots in ancient Greek philosophy, places greater emphasis on the inherent morality of actions rather than their outcomes. Also known as Duty Ethics, this theory establishes universal rules and moral boundaries that must be followed unconditionally. Kant (2011) asserts that the moral value of an action lies not in the consequences produced, but in the intention or principle (maxim) behind it. For deontologists, adherence to moral rules is paramount, regardless of the resulting consequences. Consequently, Deontological Ethics requires unconditional validity, which consists of two premises and one conclusion, and each claim is categorical. Consequently, in Deontological Ethics they tend to follow strictly the moral rules and not the positivist law, as they focus if the conclusion is moral or immoral "*An action from duty has its moral worth not in the purpose that is to be attained by it, but in the maxim according to which it is resolved*" (Kant, 2011).

Complementing these perspectives, Aristotle's (2004) Virtue Ethics shifts the focus from actions and outcomes to the moral character and virtues of the individual. This approach separates the consequence from the action and advocates for a balance between extremes, promoting ethical behavior through the cultivation of good habits and intentions. Virtue Ethics emphasizes moral development and the pursuit of a virtuous life, encouraging individuals to act with moderation and integrity.

In conclusion, while Utilitarianism, Deontology, and Virtue Ethics offer distinct approaches to ethical reasoning, each provides valuable insights into how actions, intentions, and character shape our understanding of moral responsibility. These ethical theories form the foundation for developing frameworks in fields such as Ethical AI, where the consequences associated with AI systems must be carefully considered to ensure more transparency, responsibility and fairness, which is the core of the ethical alignment.

2.2 ETHICAL AI

For instance, moral principles must be codified to address the unique challenges posed by AI, such as decision-making transparency and the potential for bias (Keng, Wiang; 2020). On the other hand, Ethical principles serve as formal guides to ensure that AI systems operate consistently with societal values, bridging the gap between individual moral intuitions and collective societal obligations. This codification of ethics into AI governance frameworks allows stakeholders to address issues that may not be immediately apparent at the moral level, such as long-term societal impacts or unintended consequences of autonomous decision-making systems (Nikolinakos, 2023).

By embedding ethical principles into the design and operation of AI systems, developers aim to ensure that these technologies align with broader societal expectations, which are essential for building trust and maintaining public confidence in AI (Zhou et al., 2020). Therefore, while morality may provide the individual compass for understanding right or wrong, ethics offers the collective map guiding AI toward ethical responsibility.

Also, many Ethical theories are being applied in the Artificial Intelligence Systems Field, such as Utilitarian Ethics, Virtue Ethics, and Deontological Ethics

(Jedličková, 2024). Each of them impacts on the development of the system differently, not only on how society carries out its responsibilities regarding AI outputs. Utilitarian Ethics may see more importance in the negative output generated by the AI system, condemning a stakeholder, even though his action did not necessarily create that result. While Deontological Ethics might focus on the stakeholder's will, if it was moral or immoral, based on good faith or not, rather them result.

The development of Artificial Intelligence (AI) technologies has given rise to numerous ethical challenges, necessitating the formulation of principles and frameworks that ensure AI systems align with moral values. It is important to highlight that Ethical AI refers to the study and practice of ensuring that AI technologies are developed, deployed, and governed in ways that promote human welfare, uphold justice, and avoid harm (Dignum, 2021). Floridi (2020) and Jobin (2019) have provided the most used frameworks for ethical AI, but their approaches differ in focus, scope, and methodology. In this section, we explore their views on ethical AI, defining and comparing their analyses.

Floridi (2020) analyzes the ethical challenges posed by AI through the lens of Bioethics. His view situates AI within the broader context of the "infosphere", defending that the digital environment is made by artificial agents and humans, therefore, he assumes that the ethical treatment of information is the core to mitigate ethical dilemmas. Therefore, he claims that AI systems must ensure that these interactions support human dignity, promote social welfare, and maintain the integrity of the infosphere. His main contribution is his ethical framework, which has four pillar principles: Transparency, Respect for Human Dignity, Accountability, and Sustainability/Social Good.

Firstly, Floridi (2020) emphasizes that AI systems operate as agent swithin the digital ecosystem. Therefore, it requires a different pattern of design, as AI systems were built focusing on enhancing and supporting harmonious human interactions. Secondly, he strongly advocates preserving human dignity and autonomy, while arguing that AI systems must augment human capacities rather than replace them or diminish human freedom. Thirdly, according to him, ethical AI systems must be accountable/responsible for their decisions, and their operations should be transparent. This is necessary to ensure that humans understand how decisions are made and can hold responsible parties accountable for outcomes. His last principle focuses on promoting not only economic progress but also long-term social and environmental sustainability. For him, AI systems should be designed to focus on the collective good, enhancing societal well-being and protecting future generations.

Jobin (2019), instead of adapting an existing ethical framework to AI, as Floridi Jobin (2019) has built the principles from the analysis of various AI ethics guidelines, as Sony (2018), The Japanese Society for Artificial Intelligence Ethical Guidelines(2017), Microsoft (2018), The Public Voice (2018) from multiple sectors and regions. Her research identified the common themes and principles across different industries, illustrating the diverse ways ethical considerations are applied to AI development. Unlike Floridi's philosophical approach, Jobin's research focuses more on the practical implementation of moral principles, acknowledging the variability in emphasis across sectors. Jobin’s core principles are Fairness, Transparency, Privacy, Accountability, and safety.

Jobin (2019) argues that the ethical guidelines for AI vary significantly depending on the context applied. Different industries and regions emphasize ethical concerns based on their specific regulatory, cultural, and operational environments. For example, for Jobin, privacy might be a top priority in health-related AI applications, while accountability might take precedence in financial technologies.

Secondly, she shares a concern about the implementation gap, advocating that there is a profound difference between the ethical principles and their practical application. While many organizations acknowledge the importance of principles like fairness and transparency, the actual process of translating these concepts into concrete actions is often underdeveloped or incomplete.

Also, Jobin (2019) and Floridi (2020) claim that AI systems must operate in ways that are just, impartial, and free from bias, emphasizing the need for AI to avoid discrimination and ensure that all individuals and groups are treated equitably. They defend the transparency principle, claiming that AI systems' operations and decision-making processes must be transparent, ensuring that stakeholders understand how AI decisions are made and can scrutinize or challenge those decisions when necessary. Besides, they defend that both developers and users of AI systems must be held accountable for the outcomes of these technologies. That includes ensuring that AI systems can be audited and that responsible parties are identified and liable for any harm caused by AI systems.

The Utilitarian and Deontological approach, Leikas et al. (2019) proposed a different point of view. Leikas et al. (2019) advocated for a multi-perspective and systematic discussion involving ethical AI values and principles. Leikas et al. (2019) advocate for a different approach to solving problems derived from AI outputs because Ethical AI would deal with different systems and incidents. For instance, moral principles must be codified to address the unique challenges posed by AI, such as decision-making transparency and the potential for bias (Keng, Wiang; 2020). On the other hand, ethical principles serve as formal guides to ensure that AI systems operate consistently with societal values, bridging the gap between individual moral intuitions and collective societal obligations. This codification of ethics into AI governance frameworks allows stakeholders to address issues that may not be immediately apparent at the moral level, such as long-term societal impacts or unintended consequences of autonomous decision-making systems (Nikolinakos, 2023).

By embedding ethical principles into the design and operation of AI systems, developers aim to ensure that these AI Systems align with broader societal expectations, which are essential for building trust and maintaining public confidence in AI (Zhou et al., 2020). Therefore, while morality may provide the individual compass for understanding right or wrong, ethics offers the collective map guiding AI toward ethical responsibility.

Also, many ethical theories are being applied in the Artificial Intelligence Systems Field, such as Utilitarian Ethics, Virtue Ethics, and Deontological Ethics (Jedličková, 2024). Each of them impacts the development of the system differently, not only on how society carries out its responsibilities regarding AI outputs. The Utilitarian Ethics approach tends to focus on blaming someone, rather than examining if the conduct was illegal or immoral. On the other hand, Deontological Ethics examines the agent's will, looking not only at the results but mainly at the agent’s good faith. Therefore, if Utilitarian Ethics is applied when a negative output is generated by AI Systems, the agent blamed for this action may not have directly acteintendd on this event. However, if Deontological Ethics is applied, negative events generated by AI Systems might be unsolved, if the society understands that the agent did not intended to generate that result.

Notwithstanding, there is a minority who claim the usage of only one ethical approach in AI Systems (Bauer, 2020; Jedlickova, 2024; da Silva, L., & Seno, E.;2023).

Our study is aligned with Leikas et al. (2019) proposal, as we understand that both Floridi’s (2020) and Jobin’s (2019) ethical principle’s frameworks complement each other. Moreover, as ethical AI deals with several occurrences that do not have precedents, it is necessary to use a combination of different ethical approaches.

Ethical principles	
Floridi et al. 2020	**Jobin et al. 2019**
Beneficence: well-being, dignity, sustainability	Beneficence, dignity, sustainability
Non-maleficence: privacy, security, and “capability caution”	Non-maleficence, privacy, technical reliability
Autonomy: the power to decide, whether to decide)	Freedom and autonomy
Justice: prosperity and solidarity	Justice, fairness, solidarity
Explicability: intelligibility and accountability	Transparency, responsibility, trust

Table 1 - Comparison of ethical principles proposed by Floridi et al. 2018 and Jobin et al. 2019 (Ximenes, Bianca, 2024)

As observed in Table 1, after examining the proposed principles of these two scholars, we noticed that they had several convergences, as the main difference between them was concerning the level of the granularity explored by each one.

Other initiatives, such as the Montreal Declaration and AI4People Ethical Framework, propose governance mechanisms to ensure the ethical development and use of AI technologies. Interestingly, frameworks like the "Google AI Principles" and the "Responsible AI Licenses (RAIL)" focus on including ethical clauses in AI licenses, offering a unique approach to integrating responsibility into the very foundation of AI deployment.

As discussed in this chapter, many of these initiatives are theoretical and do not have immediate practical applications. However, the author conducted a deeper analysis of how licenses like RAIL integrate transparency and accountability into AI system usage. This exploration highlights the potential of licenses to serve as both ethical frameworks and practical solutions for ensuring responsible AI use.

2.3 TRANSPARENCY AND CORRELATED CONCEPTS

In the context of Ethical AI, the transparency principle emerges as a fundamental and expansive concept that pertains to the ability to make the processes, decisions, and functioning of AI systems comprehensible and accessible to diverse stakeholders.

Specifically, transparency entails disclosing information regarding the data employed, the algorithms utilized, and the criteria adopted in automated decision-making (Jobin, 2019). However, it does not imply revealing every technical detail of the system. Instead, it advocates for providing an appropriate level of visibility that enables users and regulators to assess the legitimacy, fairness, and reliability of the decisions made by AI systems.

Moreover, transparency plays a critical role in addressing ethical concerns, as it is intrinsically linked to explicability, trust, and responsibility (Floridi, 2020). By fostering a deeper understanding of AI mechanisms, transparency not only promotes information about the system but also strengthens public confidence in the technology. Consequently, it serves as a cornerstone for ensuring that AI operates within ethical boundaries, balancing innovation with societal values.

Explicability is directly linked to transparency and pertains to the ability to provide clear and comprehensible justifications for the decisions or actions taken by an AI system (Cortese, 2022). While transparency establishes the foundation for access to information, explicability focuses on the communication of that information in an intelligible manner to non-technical audiences, such as end-users and decision-makers (Deutsche Telekom, 2018).

It is through explicability that the inner workings of AI are rendered intelligible and accessible, thus operationalizing the broader ethical goal of transparency (Floridi, 2018). Floridi (2019) proposes explicability as an ethical principle. However, its implementation is not without controversy. Debates often center on the practical challenges of achieving explicability without compromising technical performance or intellectual property and the extent to which such justifications meet the needs of diverse stakeholders.

The principle of explicability, as proposed by Floridi, combines intelligibility and accountability, emphasizing the need to understand how AI systems function and to identify responsibility for their decisions (Floridi, 2019). While Floridi elevates explicability as a fifth ethical principle, integrating it with autonomy, justice, beneficence, and nonmaleficence, Cortese (2022) critiques this view, arguing that explicability is better understood as an *epistemic requirement.* Rather than being an independent principle, explicability supports the implementation of the existing ethical principles by providing the clarity and oversight necessary for their fulfilment, as "*Explicability alone does not directly or per se affirm moral obligations that must always* *be acted upon"* (Cortese, 2022). Explicability, while frequently cited as an ethical principle in AI, can be more accurately understood as a tool for achieving transparency rather than a standalone principle, *"It is especially important that AI be explicable, as explicability is a critical tool to build public trust in, and understanding of, the technology*" (Floridi, 2018). Therefore, we understand it as being a tool for achieving transparency, as defended by Gilpin et al. (2018); Kuang (2017), and Wachter et al. (2017). Also, little progress has been made in developing an explicable System as defended by Robbins (2019) "*One of the main reasons that AI, and ML specifically, is the target in calls for a principle of explicability is that these algorithms are opaque".*

Moreover, Floridi (2020) emphasizes the intrinsic link between transparency and explicability:

> This principle is expressed using different terms: "transparency" in Asilomar; "accountability" in EGE; both "transparency" and "accountability" in IEEE; "intelligibility" in AIUK; and as "understandable and interpretable" for the Partnership. Though described in different ways, each of these principles captures something seemingly novel about AI: that its workings are often invisible or unintelligible to all but (at best) the most expert observers."

While transparency is the overarching goal, explicability serves as the instrument to achieve it. The ability to explain AI decisions in an understandable way is crucial for building and maintaining public trust in these systems. Without explicability, transparency cannot be fully realized, and ethical alignment becomes more difficult to ensure.

Despite these challenges, explicability remains a pivotal component of ethical AI (Cortese, 2022), bridging the gap between complex systems and societal understanding (Gilpin et al.; 2018), as can be seen below:

> *"Finally, Floridi et al. call for a principle of 'explicability' for AI which claims that when systems are powered by AI (...) A principle of explicability, then, is a moral principle that should help bring us closer to acceptable uses of algorithms." (Robbins, S. 2019)*

Through explicability, AI systems translate complex algorithms and processes into information comprehensible to non-expert stakeholders. This makes the broader principle of transparency actionable, empowering stakeholders to assess and engage critically with AI systems.

Transparency and explicability are critical to ensuring that automated decisions can be audited and understood, which, in turn, reinforces Trust. The trust principle represents the perception of users and society that AI behaves in a predictable, fair, and ethical manner. Trust builds over time through consistent interactions with transparent and explainable systems and is indispensable for the societal acceptance of AI technologies.

The transparency principle reinforces trust as it makes their internal processes and decision-making mechanisms accessible, transparency fosters trust among stakeholders. By allowing users, developers, and regulators to understand how AI decisions are made, transparency reduces hidden biases and promotes the perception of fairness and reliability in AI outcomes. Consequently, stakeholders are more likely to engage with and rely on AI systems that they view as transparent and open, promoting Trust.

Trust is a fundamental pillar that ensures AI systems are ethical in their deployment and operation. According to Gómez (2024), transparency is an essential principle, because it establishes trust among users, developers, and the public by providing insight into AI systems' internal workings, thus ensuring that decisions can be scrutinized. Also, Nikolinakos (2023) claims that trust in AI is cultivated when stakeholders can assess and evaluate AI systems with full knowledge of their design and purpose, reducing the likelihood of unexpected outcomes or harm.

Furthermore, accountability serves as the principle that ensures individuals or organizations can be held responsible for the consequences of decisions made by AI systems (Robbins, 2019). Accountability is inherently dependent on transparency and explicability, as without access to clear information and an understanding of automated decisions, it would be impossible to identify errors or assign responsibility. This principle requires the establishment of legal, technical, and institutional mechanisms to address potential harms caused by AI systems, ensuring the capacity to rectify injustices, identify biases, or address improper behaviour (Jobin, 2019). Thus, transparency, as the overarching concept, enables explicability, strengthens trust, and creates the necessary conditions for accountability in AI systems.

Also, responsibility and accountability principles are closely tied to transparency. Transparent AI practices clarify the roles and duties of each stakeholder allowing them to fulfil and articulate their responsibilities more effectively. By ensuring a clear understanding of who is responsible for specific outcomes or decisions within the AI system, transparency also strengthens accountability (Dignum, 2020).

Together, transparency, supported by explicability, enhances trust, responsibility, and accountability, forming a cohesive ethical framework that promotes responsible and reliable AI integration into society.

Moreover, transparency fosters responsibility, which is a derived aspect of trust, as it ensures that the stakeholders will be held liable when adverse outcomes occur. Therefore, we understand that the transparency principle is underpinned by trust and explicability, as they are tools to apply transparency in real-world situations.

Responsibility often refers to duty, authority, or control over something or someone (Kant, 2011). Generally, having a person responsible for the event is a social, juridical, and technical necessity (Rakova et al., 2021).

One ethical conundrum surrounding Ethical AI is the generation of negative outputs by AI systems. Typically, Negative Outputs are considered non-intended results that negatively affect the stakeholders, leading to a social or legal problem (Dignum, 2020). They can be generated by many possibilities, such as a user input error, a system bug, and a moral dilemma (Dignum, 2019). The main difficulty in this field is to indicate who is responsible for the event, as Artificial Intelligence Systems do not have the legal capacity to be responsible for it, as affirmed by Taylor (2024): "*attributing responsibility to the system itself is irrelevant, as it cannot be blamed or a subject of punishment."*

Therefore, Ethical AI aims to build systems aligned with ethical principles, looking forward to clarifying the responsibilities of each stakeholder. This only can be achieved through transparency. Also, dealing with legal responsibility is tremendously tricky due to the sanctions and powers behind it.

The main difference between “accountability” and “responsibility” lies in different comprehensions of the law. To understand that we may observe the meaning of the juridical terms and where they were forged or disseminated. Hart (1961), in 'The Concept of Law,' presents the Theory of Legal Positivism, in which rules do not necessarily connect with morality. So, a rule for *positivism* can be *amoral or immoral, but it has to be followed.* Hans Kelsen (1934), the father of legal positivism, implies that the 'positivity of law' is:

> (…) created and annulled by acts of human beings, thus being independent of morality and similar norm systems. This constitutes the difference between positive Law and natural Law, which, like morality, is deduced from a

> presumably self—evident basic norm that is considered to be the expression of the 'will of nature' or of 'pure reason.' (Kelsen, 1934)

Legal positivism implies the term ‘responsibility’ as they comprehend it as holding someone responsible for an event. Responsibility, for them, is a positive norm, an obligation, and does not necessarily have a connection to morality. Although legal positivism was created and reinforced in Germany, it was highly rejected after the Second World War, as mainly II World War acts were not illegal, even though immoral. In response to this disconnection between legality and morality, a new philosophical movement emerged in the post-war period: **Critical Theory**. Rooted in the Aristotelian principles of morality and ethics, Critical Theory sought a deeper justification, referred to as *Letzbegründung* (ultimate grounding). Key figures in this movement include scholars such as Habermas and Alexy, who aimed to address the limitations of legal positivism by integrating ethical considerations into legal and social frameworks.

In the Critical Theory vision, it started to observe that the Law is coercive and prohibitive, imposing some unwanted conduct while preventing other events that could have a positive value. These theorists seek to explain how legal legitimacy can be achieved or demonstrated, or at the least what the conditions are for a legitimate legal system (Klatt, 2007).

To mitigate legal but immoral actions, Germans focused on the ulterior motive of the action and the action itself (Klatt, 2007), looking forward to preventing the result and the norms from being ethically accepted. Consequently, they cultivated the term “accountability” differing from responsibility as the former focus an amicable and non-normative rule, based on moral principles, and the latter is a legal obligation and imposition.

2.5 FINAL REMARKS

This Chapter has provided the necessary definitions and concepts relevant to understanding the research, particularly about Ethics, Ethical AI, Responsibility, and Transparency. It began by discussing the various ethical theories, such as Utilitarianism, Deontological Ethics, and Virtue Ethics, and their application to AI systems. A multi-perspective approach was advocated for Ethical AI, combining these ethical frameworks to address complex issues in the field. The concept of Responsible AI was then introduced, focusing on moral and legal responsibilities for AI outputs, especially in cases of adverse outcomes. Lastly, the principle of transparency was explored, highlighting the importance of transparent, understandable decision-making processes within AI systems to maintain public trust and ethical alignment. These foundational concepts are critical for framing the ethical and practical challenges this research aims to address.

3. RESEARCH METHOD

This Chapter presents the methodology adopted to conduct this research. Specifically, it provides an overview of Design Science Research Methodology (DSRM) and discusses its importance in crafting innovative artefacts to solve identified problems. Also, we present the steps of the DSR cycle, including problem identification, artefact design, and demonstration.

3.1 DESIGN SCIENCE RESEARCH METHODOLOGY

Design Science Research (DSR) is a methodology for design and investigation that enables the creation of innovative artefacts aimed at solving identified problems while contributing to both practical and theoretical knowledge (Hevner et al., 2004). This approach not only explores new solution alternatives but also explains the exploratory process itself, enhancing overall problem-solving capabilities (Holmström et al., 2009). Since DSR focuses on addressing real-world problems through the development of rigorously validated artefacts, we adopted this methodology to create an artefact that fosters transparency in AI systems.

In this dissertation, we adopted Design Science Research Methodology (DSRM) due to its suitability in addressing complex challenges and developing new artefacts. We aimed to develop an innovative artefact to clarify stakeholder responsibilities while ensuring ethical alignment, therefore, DSRM provided a robust and structured approach that enabled the design, development, and refinement of solutions (Peffers et al., 2007). Also, we selected DSR because it offers a structured framework for designing innovative, practical, and theoretically grounded solutions, making it particularly well-suited for addressing complex issues such as Ethical AI (Hevner et al., 2004).

Our insights into the problem were derived from judicial cases involving negative outcomes generated by AI systems, as well as from the literature review. Based on these findings, we recognized that the documentation of AI systems is often non-transparent and difficult to comprehend, underscoring the necessity to design a new artefact with a more user-friendly layout. Figure 1 shows the phases adopted to conduct this study.

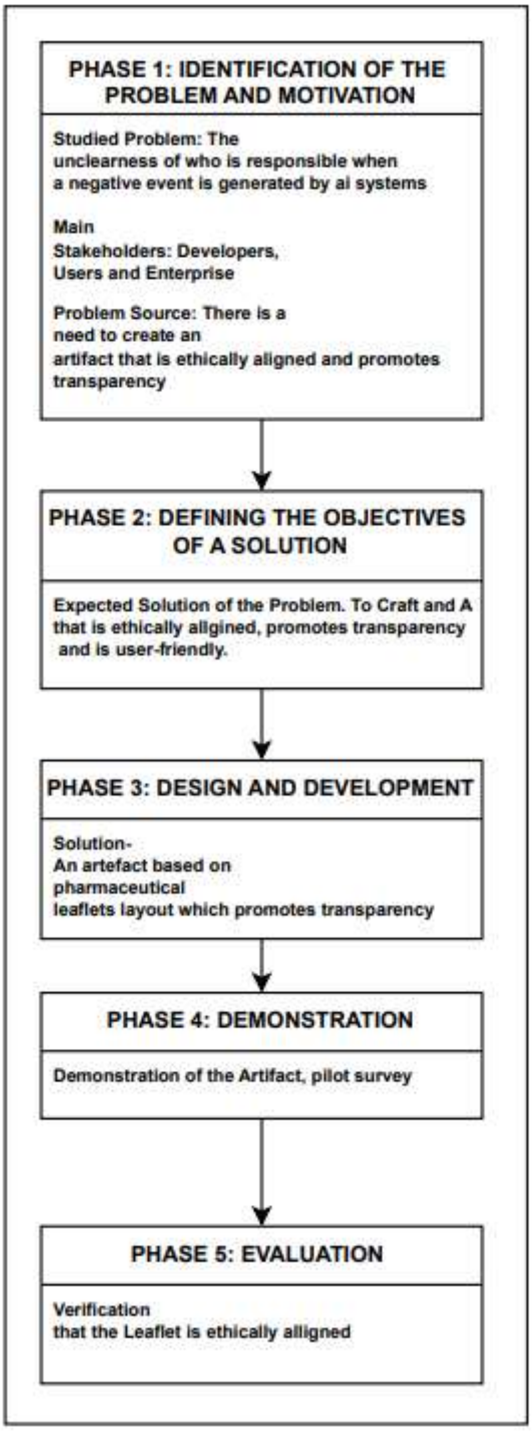
PHASE 1: IDENTIFICATION OF THE PROBLEM AND MOTIVATION
Studied Problem: The unclearness of who is responsible when a negative event is generated by ai systems
Main Stakeholders: Developers, Users and Enterprise
Problem Source: There is a need to create an artifact that is ethically aligned and promotes transparency
PHASE 2: DEFINING THE OBJECTIVES OF A SOLUTION
Expected Solution of the Problem. To Craft and A that is ethically alligined, promotes transparency and is user-friendly.
PHASE 3: DESIGN AND DEVELOPMENT
Solution- An artefact based on pharmaceutical leaflets layout which promotes transparency
PHASE 4: DEMONSTRATION
Demonstration of the Artifact, pilot survey
PHASE 5: EVALUATION
Verification that the Leaflet is ethically alligned

Figure 1: The Author, 2024

Therefore, in the *first phase* of our research - *Identification of the Problem and Motivation* - we conducted a literature review, analyzing guidelines in the Ethical AI field. We observed that these guidelines were vague and did not clearly define the responsibilities of the various stakeholders involved. Subsequently, we examined articles on Ethical AI, identifying the most cited principles. Among these, transparency emerged as a central theme.

Then, we critically analyzed the Terms of Use, the current document used in AI Systems documentation, comparing it to the ethical principles proposed by Floridi (2019) and Jobin (2019). Our analysis revealed that this document lacked clarity, particularly in defining stakeholder responsibilities. This lack of clarity was attributed to both the document's layout and its length (Dignum, 2019). As a result, we concluded the first phase of our study with the idea of developing a new artefact designed to delineate stakeholder responsibilities in cases where AI systems produce negative outcomes.

The *second phase* of our study - *Defining the Objectives of a Solution* - focused on defining the research objectives. We established that the primary goal of the Leaflet for Artificial Intelligence Systems (LAIS) was to promote transparency through an innovative and ethically grounded document that would explicitly define the responsibilities of stakeholders in AI systems. To achieve this, we conducted a preliminary analysis to identify the most effective layout to enhance transparency.

In the *third phase - Design and Development* - we designed and developed the artefact LAIS based on established principles of ethically responsible AI. The artefact was designed regarding responsible AI guidelines and validated by experts in the field.

Through our investigation, we identified significant gaps in stakeholders' responsibilities when AI systems produced negative outcomes, and we also recognized issues with the layout of existing documents. Terms of Use documents are often lengthy and difficult to understand (Dignum, 2019), causing users to overlook critical points about AI Systems. Therefore, it became essential to redesign the document layout. We crafted the LAIS with three key requirements: it must be concise, comprehensive, and transparent. As a layout model, we selected the Pharmaceutical Leaflet, which is globally recognized for its precision and user-friendliness.

Following our *ad hoc* review, we chose the Pharmaceutical Leaflets as the inspiration for the LAIS artefact. Given the objective of creating a user-friendly and ethically aligned document, the Pharmaceutical Leaflets are widely accepted, and easily understood, which made them a suitable choice. However, we needed to make some necessary adjustments to tailor the design for AI systems specifically. By adhering to the Design Science Research Methodology (DSRM), we ensured that the artefact was not only innovative but also rigorously tested, to improve how AI systems are governed and utilized in Society.

We mapped various document layouts, including Software Licenses, Terms of Use, FDA guidelines, ANVISA regulations, AI Act, LGPD, GDPR, and other ethical AI documents—looking forward to crafting the most suitable sections for the LAIS.

In the *Design and Development Phase* of this project, we mapped various document layouts—including Software Licenses, Terms of Use, FDA guidelines, ANVISA regulations, the AI Act, LGPD, GDPR, and other ethical documents to develop the LAIS.

For the *Demonstration Phase*, we conducted a survey with 18 participants to evaluate the LAIS. The participants provided both quantitative and qualitative feedback, as the survey contained four open-ended questions. The demonstration phase did not fully adhere to Peffers' et al. (2007) iterative process, as we did not develop a new version of the artefact. Given that we had already conducted a preliminary evaluation through pilot testing and incorporated additional items, we opted to retain the current version. The high level of acceptance among the participants further justified this decision.

The *Evaluation Phase* of the artefact was conducted through a survey targeted at technology professionals and users, selected based on their expertise and specialization in the field. This survey had a qualitative nature, relying on the opinions of experts in the AI domain, it was divided into three parts.

1. Identification and Participant Profile
2. Quantitative evaluation of the artefact
3. Qualitative evaluation of the artefact in general

The survey was sent to the participants with a preliminary section that explained the context, objectives, and fictional scenarios. For the first part of the survey, questions were asked about the participants' profiles, dividing them into 3 categories: a) developers b) researchers c) exclusively users. Secondly, the survey had 18 questions about each item of the proposed artefact, the participants needed to evaluate the LAIS following the Likert scale. The third part of the survey contained open questions, where the participants could justify and complement their answers.

3.2 FINAL REMARKS

This Chapter has provided a comprehensive overview of Design Science Research Methodology (DSRM) for creating innovative artefacts that address identified problems while contributing to practical and theoretical knowledge. It explained that DSRM is an iterative process involving design, evaluation, and refinement to enhance problem-solving capabilities. Also, it discussed the research phases, while demonstrating in Figure 1 the methodology of this research.

In Chapter 4, we will discuss the problem characterization, trying to understand the literature gaps. We will review the literature on how ethical principles correlate with the responsible AI field. After, in chapter 5, the author will analyze other works, looking forward to finding the state of the art surrounding responsible AI documents. We will map other documents and analyze if they lack something. In Chapter 6, the author will propose a solution, describing how it was built and done and the methods used to validate it. Later, in chapter 7, we will discuss which methods were used to validate the artefact and which protocols were utilized, with the aim to mitigating possible bias. The final chapter will present the study's conclusions and describe its contribution to the academy and industry. We finalize the dissertation, discussing possible future works.

4. PROBLEM CHARACTERIZATION

This chapter addresses the main problems related to the Ethical AI field. While studying this field, we noticed that there are two primary issues related to Ethical AI, firstly: a) the lack of transparency in AI documents, and b) the AI systems responsibility gap. The former issue might be solved by developing a user-friendly document layout, widely accepted by society and committees. For the latter, the AI documentation must be precise and objective, embedding not only accountability but also transparency.

In addressing the issue of unclearness in AI systems' stakeholder responsibilities, we prioritized the principle of transparency over other ethical AI principles, as it is critical to internal processes and decision-making mechanisms. Our choice to focus on the transparency principle was due to the author's background in law. Even though there are many ethical principles, we could not study all of them as our time and research was limited. Therefore, we chose to address the transparency principle, as unlike other principles, transparency directly addresses the need for accessible and clear documentation, which mitigates ambiguity and ensures that accountability can be upheld in cases of negative outcomes, also, it provides stakeholders to understand, assess, and assign responsibility accurately.

The research problem addresses the lack of transparency in AI documentation, which leads to ambiguity in defining stakeholder responsibilities when negative events occur. Also, the absence of a clear accountability framework creates a gap, as there is no definitive chain of responsibility among the stakeholders involved. Additionally, we examined juridical cases involving AI incidents, such as Tesla's (2021) case, and found that these legal processes are often prolonged due to the lack of clear definitions regarding developer and user responsibilities. Users have also expressed concerns about their limited understanding of the potential risks and limitations of AI systems.

4.1 THE PROBLEM OF LACKING TRANSPARENCY

The principle of transparency, as outlined by Floridi (2020) and Jobin (2019), ensures that AI systems are understandable and open to scrutiny by all stakeholders, including developers, users, and regulators (Mikalef et al., 2022).

When the principle of transparency is absent, several critical issues emerge, undermining the trust and reliability necessary for the widespread adoption and safe utilization of AI technologies.

Transparency involves making the processes and decisions of AI systems accessible and understandable, ensuring that stakeholders can evaluate their fairness and reliability (Cortese, 2022). As defended by Floridi (2019), the principle of transparency is interconnected to other principles, such as explicability, trust and responsibility.

> (Transparency) It complements the other four principles: for AI to be beneficence and nonmaleficence, we must be able to understand the good or harm it is actually doing to society, and in which ways; for AI to promote and not constrain human autonomy, our "decision about who should decide" must be informed by knowledge of how AI would act instead of us; and for AI to be just, we must ensure that the technology— or, more accurately, the people and organisations developing and deploying it—are held accountable in the event of a negative outcome, which would require in turn some understanding of why this outcome arose (Floridi, 2019).

Also, transparency not only encompasses explicability but also encompasses trust. Transparency fosters trust by enabling stakeholders to understand and assess AI decisions. Cortese (2022) defends that Trust is a byproduct of ethical AI, not a direct principle or condition, as its focus is to evaluate the integrity of AI Systems over time (MI Garage, 2019):

> "Suggestions for building or sustaining trust include education, reliability, accountability, processes to monitor and evaluate the integrity of AI systems over time and tools and techniques ensuring compliance with norms and standards. Whereas some guidelines require AI to be transparent, understandable, or explainable in order to build **trust,** another one explicitly suggests that, instead of demanding understandability, it should be ensured that AI fulfils public expectations (Jobin, pg 7, 2019)"

Additionally, the absence of transparency can lead to negative outcomes in AI systems, as stakeholders often face challenges in identifying who should be held accountable. Therefore, transparency facilitates the tracing of AI systems' decision-making processes, making it easier to identify the root causes of errors or biases.

One prominent example of the consequences of a lack of transparency is the COMPAS algorithm, used in the United States to assess the risk of recidivism. The algorithm has been widely criticized for disproportionately predicting higher recidivism risks for Black defendants, raising significant concerns regarding fairness and justice. Without transparency, biases like these remain hidden and unaddressed, perpetuating systemic inequalities within the legal system.

Another instance of transparency issues can be observed in the healthcare sector, where AI is increasingly employed for diagnostic purposes. Research has shown that an AI system designed to detect pneumonia from chest X-rays was trained on data that disproportionately represented certain demographics, leading to less accurate diagnoses for women and minorities (Lenharo, 2024). The lack of transparency regarding the training data and decision-making processes of the algorithm made it difficult for healthcare providers to understand the source of these disparities or to implement corrective measures. Consequently, some patients received incorrect diagnoses or inappropriate treatments, compromising the effectiveness and safety of the AI system (Lenharo, 2024).

Similarly, algorithms that determine content visibility on platforms such as Facebook or YouTube are often opaque, generating concerns about the spread of misinformation and the creation of echo chambers, a metaphorical description of information typically amplified by the community. Users typically remain unaware of how these algorithms prioritize content, which can amplify harmful or misleading information. This lack of transparency became particularly evident during the 2016 U.S. elections when social media algorithms were found to contribute to the dissemination of false information and the polarization of public opinion (Badawy, Ferrara, and Lerman, 2018). The opaque nature of these algorithms complicates efforts to regulate platforms and ensure that they act in the public interest. This opacity undermines the integrity of public discourse and poses a challenge to maintaining a healthy democratic process, where informed and transparent communication is essential (Lasisi, 2018).

This unclearness is problematic because while AI systems offer substantial advancements, they also introduce inherent risks due to their complexity and unpredictability. When harm results from an AI system, the public naturally seeks to hold someone accountable, particularly when civilian lives are at stake (Tigard, 2020). However, current guidelines often fail to specify who should bear responsibility, creating a "responsibility gap" that leaves victims without a clear path to justice and undermines public trust in AI technologies.

It is important to note that humans have a natural tendency to assign blame when adverse events occur (Matthias, 2004), a tendency that is heightened in cases involving vulnerable groups such as children, the elderly, and women, or when criminal activities are involved. Given that autonomous vehicle accidents and the misuse of AI in healthcare can lead to severe injury or death, it is essential to have clear documentation outlining stakeholder responsibilities.

Implementing the transparency principle in AI system’s documentation is a significant challenge due to the inherent complexity and opacity of AI technologies. Transparency requires that AI systems provide clear, accessible, and comprehensive information about their design, functionality, decision-making processes, and potential risks (Floridi, 2020). However, as Dignum (2021) highlights, many AI systems operate as "black boxes," making it difficult for even their developers to fully understand or explain how decisions are made. The lack of standardized frameworks for transparent AI documentation further exacerbates this difficulty, as stakeholders often lack clear guidelines on how to present technical information in a way that is both ethical and practical (Cortese, 2022)

Thus, the lack of transparency in Ethical AI is a profound issue, as it must be embedded as a core principle in the development, deployment, and governance of AI systems. This principle might ensure that AI technologies are not only innovative but also ethical, reliable, and beneficial to society.

The transparency principle in AI systems should be applied addressing the needs of a broad target audience, including developers, users, regulators, and the public. Each of these groups has distinct requirements for understanding AI systems. Developers benefit from transparency as it enables them to diagnose issues, improve system performance, and ensure ethical compliance in line with industry standards. Users, on the other hand, rely on transparency to understand how AI systems operate, enabling informed decision-making and fostering trust in the system's outputs Regulators require transparency to ensure that AI systems comply with legal and ethical standards, particularly concerning safety, accountability, and privacy protections. Transparency, therefore, must be tailored to these diverse audiences through clear, accessible, and context-specific documentation, ensuring that all stakeholders can engage meaningfully with AI systems, regardless of their technical expertise.

At this stage of the research, we noticed that the related works did not focus on crafting artefacts for the ethical alignment of AI systems (further discussion on Section 5.1 and Section 5.2), nor on making more transparent the content and the risks involved in these systems for the users. Also, the licenses used in this field were mainly descriptive, not being ethically aligned, as they were not transparent, responsible nor fair. Therefore, we recognized the necessity of creating a novel solution (Peffers et al., 2007; Wieringa, 2014), and it should be concise and user-friendly, ensuring ease of understanding and accessibility for all stakeholders.

4.2 FINAL REMARKS

This chapter has explored the critical issues surrounding Ethical AI. It provided real-world examples from various fields, such as criminal justice and healthcare, to demonstrate the significant impact of opaque AI decision-making and the lack of clear responsibility for adverse outcomes. Additionally, the chapter emphasized the necessity of creating a user-friendly, ethically responsible, and broadly accepted solution, aligning with both public expectations and rigorous ethical standards.

5. STATE OF THE ART

This chapter discusses the main features and limitations of proposals to use leaflet layouts to present the documentation of systems. Firstly, we analyzed the studies that proposed leaflets as systems documentation, then, we analyzed their research objective and compared their results, examining whether they promote an ethical alignment for AI Systems.

5.1 THE SOFTWARE LEAFLETS

Previous research has explored the concept of adapting the pharmaceutical leaflet model to software documentation, as exemplified by the works of Lima (2015) and Leal et al. (2012). While these initiatives were inspired by pharmaceutical leaflets, their focus was limited to improving software transparency by listing functional and non-functional requirements in an accessible format.

The earlier models did not attempt to align their frameworks with ethical guidelines or principles, focusing solely on technical and usability aspects. Furthermore, these models drew inspiration from a single pharmaceutical regulatory framework, such as the Brazilian National Health Surveillance Agency (ANVISA).

5.1.1 Lima (2015)

Lima (2015) proposed a "Software Leaflet," inspired by the format of Brazilian pharmaceutical leaflets (ANVISA). One of the main contributions of this research was the development of a Software Leaflet that presents both functional and non-functional information about software in an accessible format, as their goal was to promote understandability. The study utilized mind maps to enhance clarity and user understanding, and it emphasized the need for transparency in software development and usage. According to Lima (2015), adopting transparency "fosters better communication between developers and users by providing essential technical and functional information".

Although the Software Leaflet proposed by them addresses mainly intelligibility, their layout was a mental map rather than the Anvisa's layout of the Pharmaceutical Leaflets. Anvisa's layout follows a structured and strict format of documentation. As their research was an experimental study, their proposal's goal was to prove that the Leaflet layout helped to transmit knowledge more easily to users and developers (Lima, 2015). Also, their objective differs from ours as we focus mainly on ethical alignment.

The evaluation of the Software Leaflet was conducted through an experimental study involving 326 participants, using a survey to assess the leaflet’s effectiveness. Lima’s research aimed to answer the question: “*Is it possible to create a leaflet model for software?” and had the goal of “transmitting knowledge to those who observe the leaflet*” (Lima, 2015, p. 52). The experiment's results demonstrated that the leaflet model improved the explicability of the software, thus contributing to greater transparency in software documentation.

5.1.2 Leal et al. (2012) - BuS (Leaflet for Software)

Leal et al. (2012) introduced another Software Leaflet model, referred to as BuS (Leaflet for Software), which was also inspired by the structure and clarity of pharmaceutical leaflets. The primary aim of their research was to enhance software transparency by providing users and developers with well-organized, accessible, and standardized information on the functional and non-functional aspects of software.

One of the main contributions of Leal et al.'s (2012) research was the introduction of a hierarchical, XML-based structure for documenting software. The Software Leaflet aimed to improve usability by offering both technical and functional clarity, similar to the way pharmaceutical leaflets provide clear information to patients. While it organizes the software information into distinct sections, such as installation, functionality, and technical specifications for requirements, the authors argued that the BuS model would serve as a valuable tool for all stakeholders involved in software development and use.

The BuS model was systematically structured to facilitate communication of the software’s functional and technical details, thereby improving the accessibility and usability of software documentation. In an era where trust and clear communication are paramount, especially in global and cross-cultural contexts, the BuS model sought to make software information more understandable and reliable.

Their proposal was a new type of software documentation based on a well-structured software architecture.

> As a metaphor for the medicine leaflet, BuS was created with the aim of clarify important points about the software, with the aim of being a source of technical consultation and being able to strongly influence the transparency requirement (Leal et al. 2012)

Their artefact is not suitable for users but focuses on the format of systems documentation, based on a new software architecture. The findings of the study demonstrated that the BuS promotes transparency by detailing not only the functional requirements of the software but also the non-functional aspects, explaining how the software works, and justifying design decisions. This aligns with Leite's (2008) concept of software transparency, which emphasizes the importance of clear communication about software functionalities and decision-making processes.

5.2 SOFTWARE LICENSES

Software licenses, such as the Massachusetts Institute of Technology License (MIT), Apache License 2.0, GNU General Public License (GPL), and Berkeley Software Distribution License (BSD), play crucial roles in the legal framework governing the use and development of AI systems. These licenses grant users the right to use, modify, and distribute software and data. However, they are often criticized for their length, complexity, and lack of ethical guidelines.

The MIT License is one of the most permissive software licenses, providing users with considerable freedom to use, modify, distribute, and even sublicense software. However, from an ethical standpoint, the MIT License lacks key provisions related to responsibility and ethical use. While it includes a disclaimer that the software is provided "as-is," this disclaimer fails to address the potential for unethical applications, such as the misuse of AI technologies in areas like surveillance, weaponization, or discriminatory practices.

By not clarifying the system's limits and a responsibility chain, the MIT License places the burden of responsibility entirely on the user. This can result in the development of harmful applications without any oversight or accountability from the original developer. Furthermore, the legal protection it offers developers in terms of liability is extensive, potentially leaving users without recourse if the software causes harm or is used irresponsibly. The license does not provide any mechanism to ensure compliance with data protection laws, nor does it cover obligations regarding ethical AI use, creating potential legal and ethical gaps.

Similarly, the Apache License 2.0 is more comprehensive than the MIT License, particularly because it includes provisions related to patent rights, preventing contributors from pursuing patent litigation against users. However, like the MIT License, the Apache License 2.0 lacks stipulations regarding the ethical use of software. While it ensures legal clarity concerning patent rights, it does not address the misuse of AI technologies or the responsibility of contributors when the software is used for harmful purposes. Despite offering stronger patent protection, it still lacks transparency and specific ethical guidelines for AI use, leading to potential liability issues.

The GPL License, on the other hand, strongly emphasizes keeping software free and open source, ensuring that all modified versions remain accessible for public use. This copyleft provision is ethically significant, as it promotes openness and collaboration, aligning with certain ethical principles in AI development, such as the principle of transparency. However, the GPL focuses primarily on software freedom and redistribution rights, without addressing the moral or ethical implications of the software's outputs. It imposes no restrictions on how the software can be applied, meaning that developers of AI systems using GPL-licensed software are not bound by ethical standards regarding their technologies' deployment.

As with other licenses, the GPL does not enforce compliance with international ethical guidelines for AI. In the event of harmful outcomes or data breaches, this could lead to a lack of accountability for developers or companies that modify and deploy GPL-licensed software. Furthermore, its policy of software freedom might discourage some companies from adopting ethical frameworks, thus affecting the level of acceptance of this license in certain sectors.

Similarly, the BSD License is another permissive license, offering minimal restrictions on redistribution. Like the MIT and Apache licenses, it places the responsibility for ethical usage entirely on the user rather than the developer. A significant ethical concern with the BSD License is the absence of accountability provisions, particularly when AI systems are deployed in sensitive domains such as healthcare, education, or finance.

Furthermore, the BSD License does not include any provisions to inform users of potential risks associated with AI, such as biases in machine learning models or the possibility of harmful consequences resulting from automated decisions. The permissive nature of the BSD License, combined with its lack of ethical safeguards, could lead to legal disputes if AI systems built on BSD-licensed code cause harm, especially in the field of intellectual property. Moreover, providing specific guidance on responsible AI usage could expose developers and users to legal challenges.

In summary, while these software licenses enable significant innovation and freedom in AI development, they lack essential ethical guidelines and accountability mechanisms. Their permissive nature allows for the unrestricted use of AI technologies, which can lead to considerable moral and legal challenges, particularly when these systems are deployed in sensitive or high-stakes contexts.

5.3 FINAL REMARKS

This chapter has explored the characteristics and limitations of existing proposals for enhancing transparency in AI systems, by analyzing software licenses, Leal et al. (2012) and Lima (2015), which aimed to foster transparency by making software information more accessible, standardized, and comprehensive for users.

These findings underscore a critical gap in existing software documentation and licensing approaches: while current initiatives provide pathways to improve transparency, they do not adequately address the need for structured frameworks that clarify stakeholder responsibilities.

6. PROPOSED SOLUTION

The artefact proposed in this dissertation aims to address the transparency gap by providing an innovative, clear, and ethically aligned document. Our goal is to delineate the responsibilities of the stakeholders involved in the development and deployment of AI systems. Also, this artefact ensures that when a negative output occurs, there is a clear path to responsibility and remediation. Therefore, our solution will contribute to AI's ethical development and deployment, supporting a more responsible and ethical integration of these technologies into everyday life.

6.1 INSPIRATION

The concept of a "**Leaflet for Artificial Intelligence Systems**" originated from the observation that existing industry documents, such as Terms of Use, are often lengthy, difficult to read, and unintuitive. Following a literature review, we identified that these documents frequently omit crucial information about the risks associated with AI systems (Dignum, 2021). Consequently, we sought to develop a more transparent and clear solution. During our examination of these documents, we realized that the primary cause of their lack of clarity was related to their layout. This led us to search for a document format that would be clear, concise, and ethically approved.

In our research, we identified pharmaceutical leaflets as an ideal model, as they are globally recognized for their clarity, ease of understanding, and conciseness. Moreover, they are approved by stringent medical committees and regulatory bodies, ensuring their reliability and ethical alignment. Also, Lima (2015) and Leal et al. (2012) adopted a similar approach, basing their studies on the Brazilian medical Leaflet (Anvisa). As a result, we adopted this layout intending to provide stakeholders with a transparent and user-friendly document regarding AI systems. Our goal was to help stakeholders be well-informed about the ethical considerations and potential risks embedded in AI systems, supporting the promotion of a more responsible AI System.

After choosing the artefact layout, we compared FDA, TGA, and ANVISA leaflet models to understand what sections are found in the pharmaceutical leaflets. Table 2 presents a comparison of these proposals. These models were chosen by criteria of locality, relevance in the medical field, and design. It is essential to notice that these three leaflets studied have similar sections, as they all have crucial topics about warnings, adverse reactions, and modes of use.

While we compared these pharmaceutical Leaflets, we needed to understand the definition of each section and how we could translate them into computer science terms. Consequently, we studied the Brazilian resolution to craft a leaflet, the Resolution RDC nº. 47/2009, from ANVISA, which lists essential sections:

> Art. 4 For the purposes of this Technical Regulation, the following definitions are required:
>
> I - warnings and precautions: instructions on advance measures or warnings that favor the correct, prudent and safe use of the medicine to prevent health problems and that may indicate limitations to the use of the medicine, but which do not necessarily contraindicate it;
>
> II - leaflet: legal health document that contains technical-scientific information and guidance on medicines for their rational use;
>
> VIII - contraindication: any health condition relating to a disease, the patient or a drug interaction, which implies not using the medicine. If this condition is not observed, it could have serious harmful effects on the health of the drug user or even lead to death;
>
> XII - adverse event: any undesirable medical occurrence that occurs with a patient who has received a pharmaceutical product and that does not necessarily have an established causal relationship with this treatment. An adverse event includes any unfavorable and unintended sign (abnormal laboratory findings, for example), symptoms or illness temporarily associated with the use of the medication, whether related or not to the medication.
>
> XIII - pharmaceutical form: final state of presentation that active pharmaceutical ingredients have after one or more pharmaceutical operations carried out with the addition of appropriate excipients or without the addition of excipients, in order to facilitate their use and obtain the desired therapeutic effect, with characteristics appropriate to a given route of administration;
>
> XVII - severity of adverse reactions: refers to the outcome of a reaction after using the medication in a given patient, classified as serious and non-serious. The following situations are considered serious: death; threat to life, when there is a risk of death at the time of the event; hospitalization or extension of an existing hospitalization, characterized as hospital care requiring hospitalization or an extension of hospitalization due to an adverse event; significant or persistent disability, when there is a substantial interruption in a person's ability to carry out their normal life functions; congenital anomaly; any suspicion of transmission of an infectious agent through a medication and clinically significant event, characterized as any event resulting from the use of medication that causes the need for medical intervention, in order to avoid

> death, risk to life, significant disability or hospitalization. Any other event that is not included in the serious adverse event criteria is considered non-serious;
>
> XXI - drug interaction: is a pharmacological or clinical response caused by the interaction of drug-drug, drug-food, drug-chemical substance, drug-laboratory and non-laboratory examination, drug-medicinal plant, drug-disease whose final result may be the change in desired effects or the occurrence of adverse events;
>
> XXIII - special populations: subgroups of populations that present special characteristics, such as: children, elderly people, infants, pregnant women, diabetics, allergic to one or more components of the medication, heart disease, liver disease, chronic kidney disease, celiac disease patients, immunocompromised people, athletes and others who require special attention when using a certain medication;
>
> (...)

After mapping and analyzing the Resolution above we obtained the main sections in a pharmaceutical Leaflet: General Information, Indications of Use, Forme and Dosage, Warning and Precautions, Adverse Reactions, Drug Interactions, Use in Specific Population, Overdose, Information for the Patient. Then, we thought that it would be adequate to compare the layout of a pharmaceutical Leaflet to the Terms of Use, as the second is a document utilized for software systems with the aim of understanding which important sections must be included e in our proposed artefact.

	FDA	TGA	ANVISA	TERMS OF USE
RESUME	●	●	X	●
INDICATIONS OF USE	●	●	●	●
FORME AND DOSAGE	●	●	●	X
WARNING AND PRECAUTIONS	●	●	●	X
ADVERSE REACTIONS	●	●	●	X
DRUG INTERACTIONS	●	●	●	X
USE IN SPECIFIC POPULATION	●	●	●	X
OVERDOSE	●	●	●	X
INFORMATION FOR PATIENTS	●	●	●	X
LEGAL ADEQUACY	X	X	X	●

Table 2: Comparative of the Pharmaceutical Leaflets, Dots Means the existence and X means a absence

Then, we thought that our artefact items should be based on the FDA, TGA, Anvisa, and Terms of Use sections, but they had to be ethically aligned. Then, we included new sections and suppressed others, as presented in Table 3.

	FDA	TGA	ANVISA	TERMS OF USE	ARTIFICIAL INTELLIGENCE SYSTEM LEAFLET
RESUME	●	●	X	●	●
INDICATIONS OF USE	●	●	●	●	●
FORME AND DOSAGE	●	●	●	X	●
WARNING AND PRECAUTIONS	●	●	●	X	●
ADVERSE REACTIONS	●	●	●	X	●
DRUG INTERACTIONS	●	●	●	X	X
USE IN SPECIFIC POPULATION	●	●	●	X	●
OVERDOSE	●	●	●	X	X
INFORMATION FOR PATIENTS	●	●	●	X	●
LEGAL ADEQUACY	X	X	X	●	●
RESPONSIBILITY	X	X	X	X	●

Table 3: Comparative of the Pharmaceutical Leaflet Sections and LAIS Sections, Dots means the existence and X means absence

However, we noticed that we could not use the same terms from pharmaceutical leaflets in our leaflet for artificial intelligence systems. Consequently, we would need to change some pharmaceutical leaflets section titles, to be more adequate with computational terms. Notwithstanding, we did not change the section's title before we comprehended its definition. Therefore, we adapted the pharmaceutical leaflet's terms, to computational terms.

NOMENCLATURE IN PHARMACEUTICAL LEAFLETS	DEFINITION
RESUME	Offers a concise overview, highlighting indications, dosage, and warnings. Its purpose is to facilitate quick understanding for patients and healthcare professionals, promoting an informed treatment decision.
INDICATIONS OF USE	Defines the mode of use of the drug, based on evidence. This part is essential to guide healthcare professionals to properly prescribe the medication.
FORME AND DOSAGE	It recommended dosages for each therapeutic indication. This section provides detailed information on the appropriate administration of the medication, including instructions on the route of administration, frequency, and possible dosage adjustments.
WARNING AND PRECAUTIONS	Presents critical information about the safety and efficacy of the drug. This section also presents potential adverse events of drug interactions, specific contraindications, and necessary precautions.
ADVERSE REACTIONS	It comprises a detailed analysis of undesired responses associated with the use of the drug, documenting adverse manifestations observed during clinical trials and clinical practice. This part presents a systematic enumeration of adverse reactions, classifying them according to frequency and severity, and highlighting rare or serious events.
DRUG INTERACTIONS	Outlines possible reciprocal influences between the drug and other therapeutic substances or agents.
USE IN SPECIFIC POPULATION	Specific considerations related to the administration of the drug in certain population groups, such as the elderly, pregnant women, breastfeeding women, and patients with liver or kidney problems.
OVERDOSE	Provides potential effects and complications arising from excessive drug intake.
INFORMATION FOR THE PATIENT	An educational approach to give accessible information on various aspects related to treatment. Includes details on the therapeutic indication, dosage, possible adverse reactions, relevant precautions, guidelines for administration, and appropriate storage.

Table 4: Pharmaceutical Leaflet Sections Definitions

As we have defined the pharmaceutical leaflet sections, we needed to make some adjustments to the terms, to be technically adequate. Therefore, we started to rename the section’s title, aiming to be technically precise with computational terms, as can be seen in Table 5.

NOMENCLATURE IN PHARMACEUTICAL LEAFLET	PHARMACEUTICAL DEFINITION	NOMENCLATURE IN AI SYSTEM LEAFLET	DEFINITION IN AI SYSTEM LEAFLET
RESUME	Offers a concise overview, highlighting indications, dosage and warnings. Its purpose is to facilitate quick understanding for patients and healthcare professionals, promoting an informed treatment decision.	GENERAL INFORMATION	Summary of essential information related to the use of the AI System, as well as its functionalities

INDICATIONS OF USE	Defines the mode of use of the drug, based on evidence. This part is essential to guide healthcare professionals to properly prescribe the medication.	**INDICATIONS OF USE**	How the system should be used, its target audience, what is does.
WARNINGS AND PRECAUTIONS	Presents critical information about the safety and efficacy of the drug. This section presents potential adverse events of drug interactions, specific contraindications, and necessary precautions.	**TRUST**	Highlights scenarios in which the system may exhibit undesired behavior, inaccuracies, or adverse impacts, providing a clear understanding of the potential risks involved, as well as the bias tests already carried out.
INEXISTENT		**DATA LIFE CYCLE**	Set of steps of data from its collection to its eventual disposal. This cycle consists of several phases, including collection, storage, processing, analysis, use, sharing, and eventually, secure disposal of data.
INEXISTENT		**RESPONSIBILITY**	Information that clarifies the obligations, limitations, and responsibility of negative outputs by the artificial intelligence system. This section addresses both issues that are the responsibility of the system developer and the operator It will also point the forum to possible conflict resolutions.

Table 5: Pharmaceutical X Computational Leaflet Terms

We included some sections as the artefact needed to be ethically aligned. The **Data Lifecycle** item encompasses data collection, storage, usage, sharing, and disposal. It is essential for ensuring compliance with privacy regulations and maintaining ethical data practices. Transparency in the data lifecycle guarantees that AI systems adhere to legal frameworks such as the GDPR and align with principles of responsible data usage (Francés-Gómez, 2024). Without a clear understanding of how data is managed, users may be unaware of risks related to data breaches, misuse, or unauthorized sharing. Floridi (2020) underscores the importance of data transparency for safeguarding user autonomy by ensuring individuals are informed about how their data is handled, thus enabling them to make informed choices when interacting with AI systems.

The **Responsibility** items clarify the responsibilities of both the developer and the operator of the AI system. It defines legal responsibility, outlines conflict resolution mechanisms, and ensures that all parties are fully aware of their rights and obligations if a negative event occurs. By doing so, this item promotes ethical usage and fosters a more robust governance framework for AI systems.

These omissions were identified as significant gaps that needed to be addressed looking forward to ensuring the leaflet's comprehensiveness and effectiveness.

6.2 LEAFLET FOR ARTIFICIAL INTELLIGENCE SYSTEM (LAIS): The Artefact

The proposed artefact has five main items and 18 items. Also, the inclusion of each item within the LAIS is not arbitrary but rather a deliberate response to identified ethical gaps and practical necessities in AI systems governance. We also present the rationale for including each item elaboration looking forward to clarifying its significance within the context of ethical practices.

The artefact developed in this study addresses significant issues related to the transparency, responsibility, and ethical operation of AI systems. The LAIS aims to guide users and developers in understanding the crucial aspects of AI systems, from general information to specific responsibilities, aiming at clarifying the risks associated with such technologies.

To better illustrate the content of the LAIS, let us propose four **fictional companies** described as follows:

1. **ROUTES**: A hypothetical system for suggesting vehicle routes, emphasizing user navigation while addressing concerns about safety and reliability in route planning.
2. **IMAGE**: A fictional medical image diagnostic system, aimed at assisting healthcare professionals in interpreting clinical images while ensuring accuracy and human supervision.
3. **CREDIT**: A system designed for automated credit approval analysis, intended to streamline financial decision-making while addressing issues of bias and transparency.
4. **MUSIC**: An automated music composition system that generates original music pieces, showcasing the creative potential of AI and the ethical implications of intellectual property.

LEAFLET FOR ARTIFICIAL INTELLIGENCE SYSTEMS (LAIS)

SECTION I: GENERAL INFORMATION

1. **Name of the System**: Commercial Name of the Artificial Intelligence System. Example: ROUTES

2. **Purpose of the System:** Description of the impact of the Artificial Intelligence System on society. For example, "ROUTES aims to improve traffic in cities to offer a better quality of life for drivers".

3. **Developed by:** Name of the company developing the AI System. Example: ROUTES Inc.

The first item of the artefact tackles the fundamental issue of transparency by providing essential details about the AI system, such as its name, purpose, and developer. It clarifies its intended role and objectives within society by ensuring users have access to this basic information. This level of transparency is vital, as it helps establish trust between the user and the AI system, ensuring that all stakeholders are aware of the system's origins and the specific problem it is designed to solve. It serves to provide basic information on the AI system.

The inclusion of the item **Name of the System** provides accurate identification and a clear reference for all stakeholders. A defined name helps users, regulators, and developers recognize and distinguish the system, making communication easier and reducing confusion. In **article 13, a, of the AI Act**, it says that the system should identify the contact details of the provider.

Consistent naming also ensures alignment across documentation and standards, supporting better transparency and traceability.

The item **Purpose of the System** articulates the intended societal or operational impact of the technology. This description provides context regarding the system's goals ensuring that users and stakeholders understand its role and relevance. In Article 13, b, I of the AI Act it is seen that "The system should explain its purpose". Similarly, **Article 7(IV) of the PL de IA** mandates that systems define their finality, further supporting explicability by offering clarity about their objectives. Additionally, **Article 5 of the PL de IA** requires the provision of information about system interactions, reinforcing the principle of transparency. By highlighting its intentions, it fosters informed decision-making, allowing users to evaluate the system's alignment with societal values.

The item **Developed By** identifies the organization or entity responsible for creating the artificial intelligence system, serving for accountability and trust. In Article 3, of the AI Act, says that it should “Identify the Provider of the AI System**”. Also, Article 7(IV) of the PL of AI** mandates that systems present who developed it.

SECTION II: INSTRUCTIONS OF USE

1. System’s Utilization: Description of its main system functionalities and the recommended way to use the AI system. For example, “by sharing geolocated information from different drivers, superimposed on local maps, ROUTES suggests better routes taking traffic intoaccount.t”.

2. Target Audience: Group of people for whom the system was developed. For example, “ROUTES is a system aimed at vehicle drivers.”.

3. Recommendation of Use: Informs the need for human checking of the outputs generated by the AI system. For example, “the ROUTES system does not check whether the suggested routes pass through dangerous areas of the city, therefore, make sure that the route will not bring you unnecessary risks”

4. Level of System Autonomy: Degree of autonomy of the AI system in relation to the need for human intervention in decision-making. For example, “the IMAGE System works in “human in the loop” mode, which means that its conclusions must imperatively be validated by a medical professional”.

The second section addresses the problem of misuse and ambiguity in AI system deployment by outlining clear usage indications. This includes a description of the system's target audience, its primary functionalities, and recommendations for proper usage. This guidance ensures that the system is employed in the intended context and emphasizes the importance of human oversight, thereby mitigating the risks associated with autonomous or unsupervised use. Including these usage indications helps reduce the likelihood of harm or unintended consequences resulting from improper use.

The **Instructions of Use items** outline the AI system’s primary functionalities and recommend appropriate usage practices. This item helps to set accurate user expectations and ensure that the system is employed as intended, reducing the risk of misuse. By providing clear guidelines on usage, this item mitigates the risks associated with incorrect or unintended applications of the technology, thus promoting responsible AI deployment. Also, it is subdivided into four items: Systems Utilization, Target Audience, Recommendation of Use, and Level of System Autonomy.

The item **System's Utilization** provides a comprehensive description of the core functionalities of the artificial intelligence system and the recommended methods for its operation. This element helps to guide users on how to interact effectively with the technology, ensuring that its capabilities are leveraged as intended by its developers. In **Article 7, b, AI Act,** it says that the system should “explain of what extent to which an AI System has been used or is likely to be used”. Similarly, **Article 7(II) of the PL of AI,** underscores the importance of explicability by requiring systems to describe their recommendations and the consequences of their use. Transparency is further reinforced through **Article 7(V) of the PL of AI,** which mandates clear data categorization, and **Article 9(I) of the PL of AI**, which focuses on clarifying the finality of the data. By offering a detailed explanation of the system's features and proper usage, this item minimizes the potential for user error or unintended applications, thereby promoting the safe and efficient use of the AI system.

The **Target Audience items** identify the specific demographic or user group for which the AI system is designed. This item prevents the system from being used by individuals for whom it is not intended, thereby reducing the potential for adverse outcomes or unintended consequences. In article **13, b, IV of the Artificial Intelligence European Act (AI ACT)** dissertates that the systems should “explain the groups or persons by the system is intended to be used”. Domínguez Figaredo & Stoyanovich (2023) explains that even though there are many initiatives about Responsible AI Education, the public do not study or know about it. Therefore, it is essential to craft an artefact focusing on this group.

The absence of target audience information we identified as a significant oversight in earlier versions of the leaflet, making this item an essential addition to enhance the artefact’s clarity.

The **Recommendation of Use** provide critical guidance on the necessity of human oversight when interacting with AI systems. This item is essential to preventing overreliance on AI technologies and ensuring that users remain engaged and vigilant, particularly in contexts involving decision-making. In **Article 7, B, AI Act,** is says that the system should "*explain of the extent to which an AI System has been used or it likely to be used*". **Amershi et al. (2019)** show that recommendations guidelines are an adequate tool to attenuate usability problems. Also, **proposition 43 of the AI Act**, it says to clarify the requirements that should be applied to use the system. By emphasizing the need for human intervention, it addresses concerns surrounding AI autonomy and the risks posed by systems that operate without sufficient human supervision.

The **Level of Autonomy of the System** item informs users about the degree of autonomy the AI system possesses, specifying the extent to which human control or supervision is required. This transparency is crucial for helping users understand the AI's decision-making processes and knowing when and how to intervene. In **Article 7, g of the AI Act**, it dissertates about the extent of the output produced by AI Systems. Also, Autonomy is protected through stringent requirements for explicit consent, as per **LGPD Art. 15,** and it can be seen in **Art 8º, II PL of AI,** which infers about the Level of System Autonomy. **Verdiesen et al. (2021)** propose a framework to combine the technical, socio-technical, and governance layer to help humans comprehend better about systems Autonomy. This item responds to the growing need for greater clarity about the operational boundaries of AI systems, ensuring that users are fully aware of the level of independence their AI system exhibits.

SECTION III. RELIABILITY

1. Error Rate: Report error rates/system accuracy. Examples: "The error rate of the IMAGE system is 95%, measured by the ROC curve and AUC"; "The Template Matching method was used to evaluate the accuracy rate of the IMAGE system, which is 90%"

2. Bias Test: These are tools and methods used to test possible biases, such as gender, age, racial or any other relevant discrimination, reporting the results. For example, "In the development of the CREDIT system, a gender bias test was conducted using the GenBit algorithm, revealing that the model demonstrates an accuracy of 85% for male applicants and 75% for female applicants".

One of the critical concerns in AI system deployment is the reliability of its outcomes. To mitigate these concerns, we built the third section of the artefact, which reports the system's error rates and the results of bias tests. This section clarifies potential misuse, inaccuracy, and biased results by providing this information. This is especially important in some specific areas, such as healthcare or criminal justice, as AI biases have significant consequences. By clearly understanding the system's reliability, users can make informed decisions and exercise appropriate caution when utilizing its outputs.

The **Reliability** item is critical, as it addresses the system's performance metrics and potential biases, ensuring its trustworthiness and ethical deployment. It is divided into two items, Error Rate and Bias Tests. This information allows stakeholders to assess the system's reliability for its intended application and make informed decisions regarding its use. Together, these elements reinforce the system's credibility and support its alignment with societal and regulatory expectations, making reliability a cornerstone for responsible AI deployment.

The **Error Rate** item provides users with information about the system's accuracy and error rates. This item is critical for fostering realistic expectations regarding the AI system's performance and helping users assess the reliability of its outputs. As discussed in **Proposition 14 of the AI Act**, it is necessary to consider the scope of risks that an AI system can generate. Similarly, the **PL of AI** reinforces this through **Article 13**, which highlights the importance of addressing system risks, and **EMENDA Art. 9(I and II)**, which explores the autonomy of AI systems and their associated risks. Moreover, **Article 8(V) of the PL de IA** stresses the significance of allowing human interference, providing a safety mechanism to mitigate potential errors or harms. Building trust, as required by **Article 3(VII) of the PL of AI**, is further supported by transparent reporting of error rates and system reliability.

By being transparent about the system's limitations, this item mitigates the risks associated with overestimating the system's capabilities and emphasizes the importance of critical evaluation in AI usage.

The **Bias Testing** item ensures that the AI system has been evaluated for potential biases, a fundamental aspect of ethical AI deployment. This item promotes fairness and accountability by making bias testing results transparent to users, thereby reducing the risk of perpetuating harmful biases through AI applications**. In Article 10, f, of the AI Act**, explains the examination of possible bias. Including these items reflects the commitment to ensuring that AI systems do not inadvertently reinforce societal inequalities. Also, the data lifecycle might help to mitigate this possible biases.

SECTION IV: DATA LIFECYCLE

1. Data Collection: Explains what data the AI system collects, in addition to explicitly requiring user acceptance. Example: “All text that the user types when interacting with ChatGPT is stored by the system and can become public.”

2. Data Storage: Indicates how data is stored, for how long, the level of security, and whether it is anonymized. Example: “The CREDIT system follows ISO 27001 cybersecurity protocols when storing data. It anonymizes data and deletes sensitive data; among other LGPD requirements and ANPD guidelines. Data is stored for an unlimited time but can be deleted at the owner’s request.”

3. Scope of Use: Provides details about the types of data that will be collected (texts, images, audio, etc.), as well as the purpose of this collection. Example: “The ROUTES system collects geolocation information from drivers in real-time while using the application. Such data is used to improve the ROUTES system”.

4. Data Sharing: List with whom the collected data can be shared, and whether it can be used by other systems within the same company. Example: “The images and diagnoses collected by the IMAGE system, in addition to serving to improve the system, can be used by other systems from the same developer. However, such data will be passed on to any third party under any circumstances.”

5. Data Disposal: Informs how the user can request the modification or disposal of their data. Furthermore, it also informs the minimum time needed for the AI system to relearn without using the deleted or modified data. Example: “The company CREDIT Inc adopts data disposal practices that ensure the safe and permanent removal of user information when requested. Deletions and modifications of personal data can be made via the website www.credit.com/sac or by calling SAC at 0800 878 787. At most every 3 months, Credit's machine learning models are retrained to take into account changes in their databases.”

6. Intellectual Property: Explains that it has obtained authorization from the owners of the data used in training the AI system and indicates a path for complaints. Example:

> “All musical examples used to train the MUSIC system were either in the public domain or were duly authorized for this purpose by whoever owned their intellectual property. For any complications, please contact www.music.com/sac”.

The fourth section is about **data management**, which is a core concern in the ethical operation of AI systems, particularly about compliance and privacy. It details the data lifecycle, including data collection, storage, usage, sharing, and disposal processes. Even though many could say that this information can be found in the Privacy Politics of the AI Systems, or Terms of Use, we understand that the Leaflet for Artificial Intelligence Systems should contain this information as the cited documents are often complex and long, and the users do not read it in totality. Therefore, if we did not include this section in the LAIS, the stakeholders would not know about it, as they would not read the other documents.

Afterward, this section ensures users understand how their data is handled and provides transparency regarding the AI system's data practices. Doing so promotes ethical data management and aligns with regulatory frameworks, thereby reducing the risk of non-compliance and enhancing trust in the system.

The **Data Lifecycle items** is divided into several subitems—**Data Collection, Data Storage, Scope of Use of Data, Data Sharing, and Data Disposal**—each addressing a different aspect of data management. The **Data Collection subitems** explain what data the AI system collects and requires explicit user consent, aligning with principles of transparency and user autonomy. The **Data Storage subitems,** it describes how and for how long data is stored, ensuring transparency in data management practices. Also, the **Scope of Use of Data subitems** clarifies how collected data will be used, which is critical for building user trust and ensuring ethical data use. The **Data Sharing subitems** inform users about who may access their data, addressing privacy concerns and promoting ethical data-sharing practices. Finally, the **Data Disposal subitems** ensures that users are aware of their rights to request data modification or deletion, reinforcing commitments to data anonymization and privacy. In **Proposition 11, of the AI Act**, it says that each AI System should follow the regulation, and in **Proposition 46 of AI Ac**t it claims that the AI System should present information about the **data lifecycle** and compliance. Also, the LGPD addresses data treatment under **Art. 9(I)** and international data sharing in **Art. 33**, requiring organizations to operate transparently and respect non-maleficence principles.

According to **GDPR Art. 32**, agreement to **data usage** must be explicit, ensuring individuals comprehend the scope and purpose of their data utilization (**GDPR Art. 23**). The protection of personal data is reinforced by **GDPR Art. 26**, which prescribes data anonymization to prevent misuse**. Data sharing,** governed under **Arts. 47, 50, and 44 GDPR** requires adherence to transparency and the principle of justice, ensuring equitable data practices across jurisdictions. Also**, Loi & Spielkamp (2021)** claim that transparency in AI Systems is essential for the utilization of data in the public sector, as ethical autonomy is preserved through measures that allow individuals to control their data.

The **Intellectual Property** items clarify ownership rights regarding data and outputs generated by the AI system. As noted in **Proposition 43 of the AI Act**, it is essential to inform users about the requirements related to **intellectual property**, thereby mitigating misunderstandings. These items play a pivotal role in protecting users' intellectual property rights and establishing clear mechanisms for addressing potential disputes. The **PL of AI** supports these protections through **Article 27**, which addresses civil liability, and **Article 21(III)**, which outlines protocols for managing data and its ownership. Similarly, **LGPD Article 50** promotes governance and good practices, ensuring transparency and accountability in the management of intellectual property. These items are essential for protecting users' intellectual property rights and provide a clear mechanism for addressing potential disputes. Its inclusion ensures fairness and accountability, fostering mutual understanding between users and developers regarding intellectual property considerations.

Section V: RESPONSIBILITIES

1. Forum and Conflict Resolution: Forum and Channel for conflict resolution. Example: The forum for resolving conflicts involving the CREDIT system will be in the City of Recife, Brazil. Extrajudicial notifications must be sent to juridico@credit.com"

2. Developer and Operator Responsibility (statement): "Our commitment as developers and operators of the AI system is to ensure ethical, transparent, and responsible treatment of user data. In this way, our objective is to generate results that benefit society, minimizing possible negative impacts generated by the system, as well as seeking to mitigate any possible form of discrimination generated by AI."

The final section of the artefact clarifies **the responsibilities** of both the developer and the operator of the AI system. It defines legal responsibility, outlines conflict resolution mechanisms, and ensures that all parties are fully aware of their rights and obligations if a negative event occurs. By doing so, this section promotes ethical usage and fosters a more robust governance framework for AI systems.

The **Forum and Conflict Resolution** item provides users with a clear legal pathway for resolving disputes related to the AI system. **In article 13, a, AI Act,** it claims that the AI System should identify the contact details of the provider. Also, governance practices are emphasised in **LGPD Art. 50**, supporting transparency and beneficence. Conflict resolution mechanisms are detailed in **Art. 55**, enabling just and fair outcomes. Autonomy and accountability are reinforced through **Art. 8(II) of PL of AI,** which specifies the permissible level of system autonomy, and **Art. 8(V) PL of AI**, which allows for human oversight. Ethical autonomy is further addressed in **Art. 13 PL of AI,** which highlights the risks associated with autonomous systems. The principle of beneficence is upheld by ensuring system recommendations consider the user's well-being **Art. 7(II)**. These articles collectively ensure compliance with ethical norms, including justice and explicability. Also, this item is included to uphold principles of legal clarity and fairness, ensuring that users are informed about where and how to resolve conflicts that may arise from the AI system's use.

Lastly, the **Developer and Operator Responsibility** item delineates the ethical and legal responsibilities of the AI system's developers and operators. This item is crucial for ensuring accountability for the system's outputs and aligns with broader ethical concerns about the societal impact of AI technologies. **Proposition 76 of the AI Act** says that AI Systems should ensure that providers take into account AI Systems risks. Administrative sanctions, as prescribed by **LGPD Arts. 52 and 53**, reinforce organisational accountability. Obligations of operators and controllers, outlined in **Arts. 37–40**, ensure ethical alignment in data processing. Controller and operator responsibilities are delineated in **GDPR Arts. 51** and **24**, while administrative sanctions are outlined in **GDPR Art. 82**. These articles underscore the necessity for accountability mechanisms to promote good governance (**GDPR Art. 50**) and compliance with ethical principles. Winfield & Jirotka (2018) propose a new framework for the development of AI systems in robotics, entitled "*responsible innovation". It would "undertake ethical risk assessments of all new products, and act upon the findings of* *those assessments" (Being) "A toolkit, or method, for ethical risk assessment of robots and robotic systems exists in British Standard BS 8611"*

These measures aim to protect individuals from potential harm, adhering to the ethical principles of non-maleficence and beneficence. By clearly defining these responsibilities, the items support the need to hold developers accountable for the outcomes their AI systems produce.

By solving these problems, the artefact is a practical tool to improve the ethical alignment and responsible use of AI systems across various sectors. In conclusion, the artefact developed in this study addresses key ethical challenges in AI deployment by providing clear, structured guidance in areas critical to ensuring transparency and responsibility.

6.3 Final Remarks

In this chapter we explained how the proposed solution was built, from its inspiration until the final development of the artefact. In this section, we presented how the artefact was developed, its many phases, pilot tests, and detours. Additionally, the chapter concludes with the proposal of the artefact, the Leaflet for Artificial Intelligence Systems that has already been validated by the specialists and is ethically aligned with Floridi (2020) and Jobin (2019) principles.

7. ARTEFACT EVALUATION

This chapter describes the "Leaflet for Artificial Intelligence Systems" evaluation process. The objective is to evaluate whether the Leaflet for Artificial Intelligence Systems meets the ethical standards required for responsible AI deployment while remaining accessible and user-friendly for a broad audience.

To evaluate our solution, we applied a survey. The survey was conducted between April and July 2024, and participants came from diverse backgrounds: AI developers, researchers, and users. Given that questionnaires are widely used in computer science research, we considered this method to be an appropriate means of evaluating our artefact. Questionnaires typically consist of statements, questions, or stimulus words with structured response categories, often concluding with a rating scale (Oosterveld, 2019). Considering the need for the precise evaluation of the artefact, we conducted the survey using a Likert Scale to obtain structured and quantifiable responses. Although survey validation is a common practice in computer science, it is not without challenges. These challenges include unclear instructions, issues with construct validity, and insufficient feedback (Elangovan & Sundaravel, 2021). To address these concerns, we designed our survey following Elangovan and Sundaravel’s (2021) recommended steps. Additionally, we incorporated a qualitative section to allow for open-ended feedback, thus mitigating the risk of bias and enhancing the overall validity of the responses.

The survey consists of 26 questions designed to capture the perspectives of participants comprehensively (see appendix 1). Three questions concern about the participant identification. The other sixteen questions correspond to the evaluation of each item of the leaflet. And the last five questions were opened.

It begins with a brief introduction outlining the survey’s purpose and the estimated time for completion, which was approximately 20 minutes. Participants were then prompted to identify their roles—such as AI developers, researchers, or users—so that their answers could be appropriately contextualized.

7.1 PARTICIPANTS PROFILE

The decision to involve a limited number of participants—specifically 18 experts—to evaluate our proposal was due to the theory that small groups of domain experts can provide high-quality feedback that is both detailed and actionable (Nielsen; Landauer. 1993). Also, studies involving as few as five to twenty participants can uncover most usability issues, as diminishing returns occur with larger sample sizes (Turner et. al. 2006).

Our decision to conduct a survey with a few participants allowed us to capture a range of expert opinions while maintaining the feasibility of conducting detailed and iterative evaluations. These participants, selected based on their diversity backgrounds correlated to computer science, provided critical insights into the Leaflet's relevance, usability, and alignment with ethical guidelines. Their expertise ensured that the evaluation was rigorous and reflective of the nuanced requirements of AI systems documentation. Moreover, focusing on a select group of experts is consistent with best practices in participatory design, where the quality of feedback is often more important than the quantity of participants. This approach ensured that the evaluation of the Leaflet was informed by deep, domain-specific knowledge, providing a robust foundation for its application in real-world contexts.

More precisely, we chose the experts to provide a broad spectrum of perspectives. They were grouped into 3 (three) categories, as follows:

- Profile “users and researchers not in computing”:
    1. an expert on law, with PhD, running a research group in “law and technology” in a university;
    2. an expert in the public sector, who previously served as the Secretary of Science and Technology for a Brazilian state;
    3. an expert on sociology, with PhD, professor in the domain in a university;
    4. an expert in philosophy, with PhD, currently a professor in the domain in a university;
    5. an expert on education, with PhD, currently professor in the domain in a Brazilian university.
- Profile “developers in private organizations”:
    1. An expert in AI, with PhD, currently head of AI in a multinational company;
    2. An expert in AI, with MSc, developer in an important innovation institute;
    3. An expert in AI, with PhD in computing, currently CEO of a company on the AI domain;
    4. An expert in AI, with MSc, developer in a Governmental Institute;
    5. An expert in AI, with a Bachelor degree, a developer in a private company;
    6. PhD Student on AI, who works in a private innovation institute.
- Profile “academic researchers in computing”:

1. A researcher in AI with extensive experience in the market, who has founded a company in the AI domain;
2. A researcher in AI with extensive practical experience in generative AI;
3. An expert in law, with PhD, running a research group in “law and technology” in a university;
4. An expert in AI, with PhD, currently a professor in the domain in a university;
5. An expert in AI, a researcher with practical experience in generative AI, with PhD, currently a professor in the domain at a university:
6. An expert in AI, PhD student, currently a professor at a private institute and developer in the domain
7. A researcher in computing, with PhD, currently a professor at a university.

Before starting, participants were informed about the ethical implications of the Leaflet and encouraged to consider its potential impact on both the AI community and society at large. Questions regarding the ethical alignment and innovativeness of the Leaflet were included to gauge whether the artefact met the ethical standards expected by the AI community. This ensured that participants not only assessed the technical aspects of the Leaflet but also its broader societal implications.

Participants were asked to rate the importance and adequacy of all items proposed for the Leaflet using a Likert scale, which ranged from 1 (Totally Disagree) to 5 (Totally Agree). The sections evaluated included critical components such as descriptions of system functionalities, target audience identification, recommendations for use, classification of autonomy levels, error rates, bias testing, potential risks, data collection and storage practices, and intellectual property concerns. The primary objective was to determine whether these sections effectively addressed the ethical and practical requirements necessary for the responsible deployment of AI systems.

In addition to the structured Likert scale questions, the survey incorporated open-ended questions to allow participants to suggest additions, removals, or modifications to the Leaflet. This qualitative data was essential for gaining a deeper understanding of the nuanced concerns expressed by both experts and users. It also helped identify potential gaps that the Leaflet might need to address, thus ensuring a more comprehensive and responsive design.

Quantitative data from the Likert scale was aggregated to assess the consensus on the necessity and adequacy of each section, while qualitative feedback was thematically analyzed to extract common suggestions and concerns. This dual approach to data analysis allowed for a comprehensive evaluation, integrating both statistical trends and individual insights.

The feedback gathered from the survey played a pivotal role in refining the **Leaflet for Artificial Intelligence Systems**, ensuring that it meets the ethical standards required for responsible AI deployment while remaining accessible and user-friendly for a broad audience. The diverse responses provided valuable insights that contributed to the development of a well-rounded and ethically sound artefact, ultimately enhancing the transparency and accountability of AI systems.

7.3 SURVEY PREPARATION

We then sent two pieces of information to each participant in the study: the LAIS (as presented in Section 6.2 including the examples of the fictitious companies) and a link to a Google Forms questionnaire. The participants should read the LAIS (Appendix 1) before answering the questionnaire (Appendix 2).

. Concerning the **questionnaire** (Appendix 2), it was meticulously designed to gather both quantitative and qualitative insights regarding the proposed **Leaflet for Artificial Intelligence Systems**. It consisted of two primary sections: a series of closed-ended Likert scale questions and a set of open-ended questions aimed at eliciting detailed feedback from participants.

The **first section** of the questionnaire contained a single question for each key item in the leaflet. These questions were phrased as “*Should the Leaflet contain the item X?”* as ‘X’ being the name of the Item included in the LAIS, and participants were asked to respond using a Likert scale. This scale ranged from strong disagreement (1) to strong agreement (5), enabling the collection of structured data to measure the perceived relevance and necessity of each section of the leaflet. Consequently, this quantitative data provided a clear understanding of participant consensus on the inclusion of specific sections.

The second part of the survey included a series of open-ended questions, designed to encourage participants to share their subjective perspectives and constructive criticism about the leaflet. These questions were:

1. **What do you think about the initiative to create the Leaflet for AI Systems?**
2. **Would you remove anything from the Leaflet for AI Systems?**
3. **Would you add anything to the Leaflet for AI Systems?**

4. **Would you modify anything in the Leaflet for AI Systems?**
5. **If you have any additional comments, please include them here.**

This open-ended format allowed participants to provide nuanced and detailed feedback, ensuring that the study captured a wide range of opinions and suggestions. So, the combination of quantitative and qualitative approaches was critical for thoroughly evaluating the relevance, comprehensiveness, and usability of the proposed leaflet.

As part of the iterative development process of the Leaflet, we conducted a pilot survey to gather initial feedback on its length and understandability. This pilot survey involved two participants: an experienced AI system developer and a researcher specializing in AI. The aim was to evaluate the Leaflet's clarity, comprehensiveness, and relevance from two distinct perspectives within the AI field.

The feedback obtained from this pilot survey was instrumental in refining the Leaflet. Both participants highlighted the importance of including more detailed sections on system autonomy, error rates, and data lifecycle management—areas that were initially underdeveloped. Furthermore, their input validated the decision to use fictional software examples to illustrate the Leaflet's principles, as both participants found these examples helpful in contextualizing the information.

7.3 SURVEY RESULTS

We conducted a comprehensive survey with participants (Appendix 1) to validate the Leaflet for Artificial Intelligence Systems (LAIS). This survey was conducted between April and July 2024 and included a diverse cohort comprising AI developers, researchers, and users with varying levels of expertise. In Figure 2, we present the percentage of agreement of the specialists of each item of the Leaflet for Artificial Intelligence Systems.

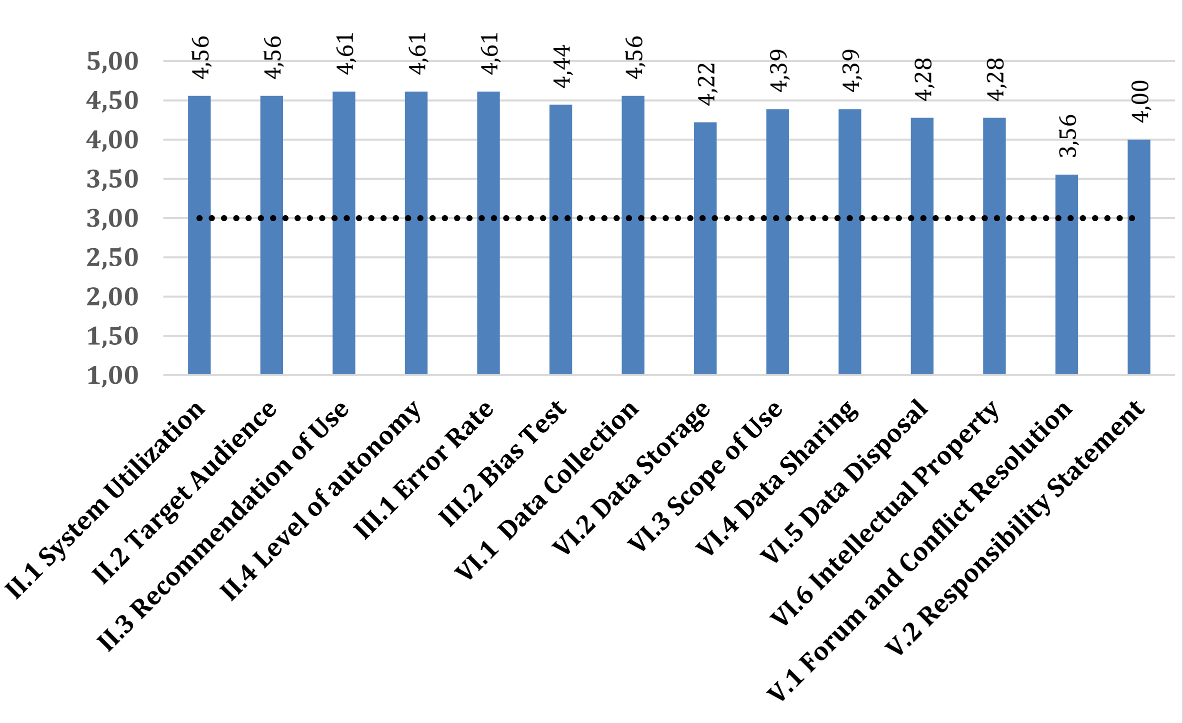

Figure 2 – Mean of the responses for a 5-point Likert scale for each item of the LAIS

The survey results revealed strong agreement among participants regarding the importance of including various sections in the Leaflet for Artificial Intelligence Systems (LAIS). Across the evaluated items, participants consistently rated the statements above 4 on a 5-point Likert scale, reflecting a broad consensus on their relevance and adequacy within the document of the sections: **System Usage**, which outlines the system's main functionalities, **Target Audience**, which specifies the intended users of the system, **Recommendations for Use** section, detailing considerations such as the target audience, duration of use, and the necessity of human oversight, **Bias Testing** section, which describes the biases tested in the system and **Autonomy Level of the System**, which describes the level of autonomy that the system have. Additional critical sections, such as **Potential Risks**, **Data Collection**, **Data Disposal** and **Scopus of Data Utilization**, also received the most favourable evaluations, even though they received less than 3 points from one or two specialists. Overall, the responses demonstrate a clear endorsement of these sections as integral components of the Leaflet.

The statement, ***"The Leaflet must contain the Error Rate section, which reports the error and accuracy rates of the system"*** received strong agreement from 13 participants. Only one participant expressed partial disagreement, assigning a rating of 2, while none of the respondents rated this section as 1 or 3. This indicates an absence of strong disagreement or neutrality. This participant commented:

> *P1: It may be that error rates, like those used in the example, do not apply to the system (e.g., systems that extract useful information from users or generate content, such as ChatGPT, may not necessarily have hit rates or accuracy indicators). (...) Furthermore, laboratory test conditions may not reflect day-to-day usage. Therefore, a disclaimer may be necessary, such as: "Hit rates are an indication," as conditions and modes of use could affect performance.*

The statement, ***"The Leaflet must contain the Data Storage section, which addresses the duration, purpose, and mode of data storage"*** received strong agreement from 10 participants. However, three participants rated this section as 2, disagreeing regarding the inclusion of the item or its content. Notably, there were no ratings of 1 or 3, suggesting that participants generally recognized the importance of this section, even if opinions varied slightly. Figure 2 shows the percentage of agreement and disagreement of the participants about the inclusion of this item in the LAIS.

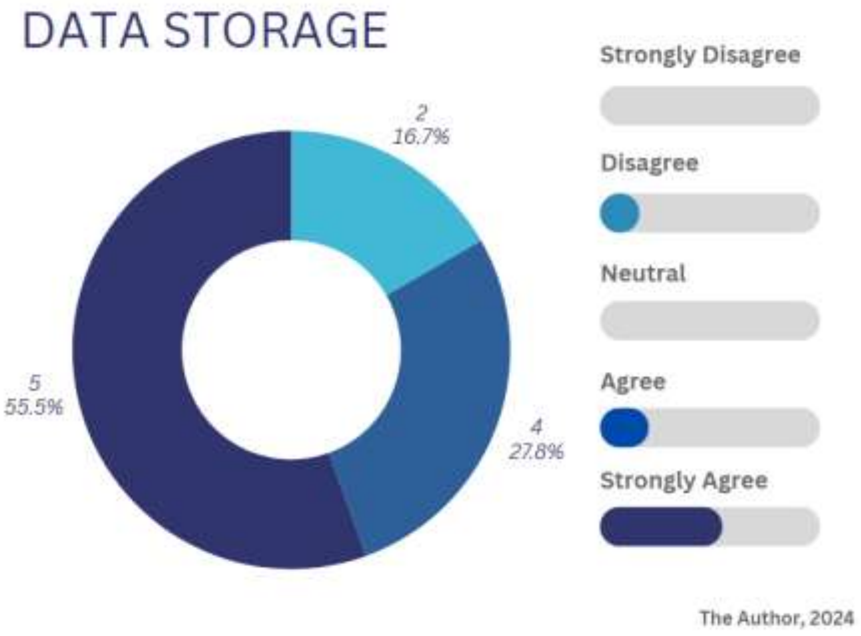

Figure 3: Data Storage Section

The qualitative responses of the participants who disagreed about the inclusion of justified by explaining that:

> **P1:** *There is a legal issue with the LGPD in Brazil, so the leaflet does not seem to be the right place to guarantee compliance.*

> **P2:** *I would make the Collection, Storage, Disposal, and all other items related to data conditional. There needs to be some kind of usage application. For example, a pre-training model differs from a network architecture or a database. Each category has specific characteristics, making a generalised leaflet problematic. In a pre-trained model, it is essential to know what data was used during training. However, for a network architecture that does not yet have the data, this section would not be necessary.*

This section's mean (average) rating was 4.17, reflecting a general consensus among participants, though slightly less unanimous compared to other sections. The median rating was 5, indicating that more than half of the respondents considered this section highly important. The variation in responses suggests that while the importance of data storage is widely acknowledged, differing views exist on how this issue should be addressed within the Leaflet.

The statement**, *"The Leaflet must contain the Data Sharing section, which lists with whom the collected data may be shared, who the system's data operators are, and whether this data may be used by other systems of the same company"*** received strong agreement from 13 participants. However, three participants rated this section as 2, indicating some disagreement or concern regarding its inclusion or content. Notably, no responses rated this section as 1 or 3. Two participants who rated the Data Storage and Data Utilisation sections negatively also rated this section as 2. Additionally, another participant rated it as 2, justifying, *"The issues related to data and its use should be reviewed, as the LGPD already addresses some of these points."*

This section's mean (average) rating was 4.28, reflecting a generally strong consensus on the importance of transparency in data-sharing practices within AI systems. The median rating was 5. However, the variation in responses indicates differing opinions on how these practices should be implemented.

The statement, ***"The Leaflet must contain the Intellectual Property section, which informs users of the channels to contact if they notice a violation of intellectual property in the database used by the AI system****"* received strong agreement from nine participants. Another six participants rated it as 4. Notably, no participants rated this section as 1, suggesting a consensus on its relevance. One participant, however, rated it as 2, justifying, ***"****Intellectual Property is already addressed in other documents, such as terms of use and the user manual. Therefore, it does not need to be included in the LAIS.****"***

The statement, ***“The Leaflet must contain the Developer and Operator Responsibility section, which declares their commitment to being ethically and legally aligned while ensuring transparency in understanding the AI system”*** received strong agreement from nine participants. However, responses revealed a diversity of opinion: three participants rated it as 4, another three rated it as 3, and another three rated it as 2. These results reflect differing views on the necessity of regulating AI responsibility. Figure 4 presents the percentage of agreement and disagreement of the participants about the inclusion of this item in the LAIS.

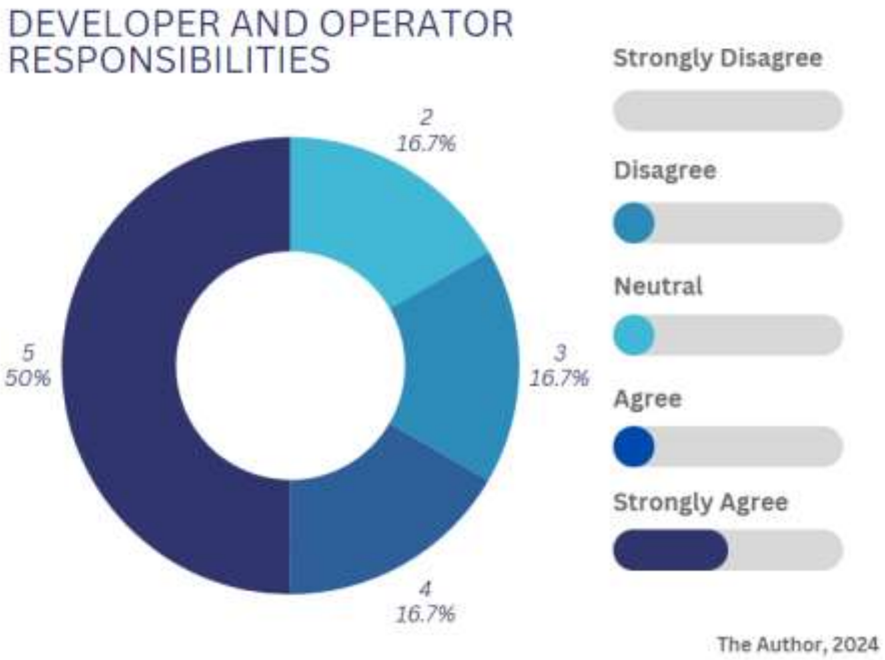

Figure 4: Developer and Operator Responsibilities Section

Three participants which disagreed with the inclusion of this item in the Leaflet justified their opinion by explaining that:

> **P1:** *“The fields are vaguely described.”*
> **P6:** *“I understand that every professional already has a commitment to and responsibility for user data (it is regulated by law). The Leaflet does not add new or relevant information about the product's functioning. This statement will not give the professional greater or lesser responsibility for user data.”*

P5 did not justify their opinion, even though rated it as 2, disagreeing with the statement. These findings suggest that while the commitment to ethical and legal transparency is recognized as necessary, the way it is communicated or enforced within the Leaflet may require further consideration or refinement to address the concerns of all stakeholders. The distribution of responses underscores the need for a balanced approach that ensures transparency while taking into account the practical and ethical implications of such commitments.

The section statement "***The Leaflet must contain the Forum and Conflict Resolution section that indicates the way and place where conflicts will be legally resolved"*** received disagreement of the participants, 5 of them rating it as 2. However, they did not justify their rate. Therefore, we maintained this section in the leaflet as it is correlated to the **Responsibilities of the Developer and Operator** section.

7.4 GENERAL OPINION ABOUT THE LEAFLET

The survey had 3 general statements about its research objective.

- *"It is pertinent that there is a Leaflet for AI Systems"*
- *"This Leaflet for AI Systems is innovative"*
- *"This Leaflet for AI Systems is ethically aligned"*

For the first statement, 15 participants strongly agreed, and 2 agreed with it. The second statement had 9 participants strongly agreeing and 9 participants agreeing with it. The last statement had 11 participants strongly agreed and 3 participants agreed.

In the qualitative question of the survey, Participant 5, rated 3 the statement ***"This Leaflet for AI Systems is ethically aligned"*** justifying that *"There are many initiatives on the topic. It would be fantastic if the Insert were something automatic, generated by the AI itself, or it will be just another initiative."*

These results indicate that the participants generally agreed with including the various sections in the leaflet, with the highest ratings observed in the sections concerning **Utilization of the System**, **Target Audience**, **Intellectual Property**, and **Developer and Operator Responsibility**. These sections were rated highly, reflecting their perceived importance in ensuring the transparency and accountability of AI systems.

The qualitative questions of the survey proved that the Leaflet for Artificial Intelligence Systems was highly approved by the specialists:

> **P1:** *I see that there are many initiatives on the topic. It would be fantastic if the Insert were something automatic, generated by the AI itself. If there is more burden for those who develop, I see it as something more distant.*

> **P2:** *I think the Leaflet idea is great. By proposing a simple and direct text (as in the medicine leaflet), it favours the user, makes it clearer what he is using and agreeing to use, in the sense of making it easier for him to understand what it is for (purpose), reliability, the origin and destination of the data (how it is collected and how user data will be used, discarded and shared) and the developer's responsibilities towards the user. In short, the idea is great. And the user, being able to read the leaflet beforehand, can agree to the terms of use and consent requested by the developer, more calmly. The terms that we "agree" on today are very problematic, as they seem to target only the developer (the company that owns the application, etc.) and not the user who will consume that program or application.*
>
> **P3:** *The LAIS could help ensure that end users better understand the capabilities, limitations and potential risks associated with the use of AI systems.*
>
> ***P4:*** *It is Very important, considering the growing use of AI for the most varied purposes and with construction and use increasingly lacking an understanding of the technology and its risks.*
>
> ***P5:*** *I believe that the leaflet is a very useful tool for sustainable AI and that it must be aligned (compliance) with other tools that already exist in Brazil and around the world.*
>
> ***P6:*** *I believe that this type of informative material is significant in bringing more direct communication with users. As mentioned in the material itself, the terms of use/system licenses are commonly extensive and verbose. This type of leaflet can synthesize the main information of AI systems, facilitating the sharing of this type of information, empowering users (allowing them to have the possibility of understanding a little more about the systems and using them critically) and helping to increase transparency and of these trust systems.*

In sum, the general opinion about the Leaflet was positive, focusing on its innovation and possibility to promote transparency in AI Systems.

7.5 FINAL REMARKS

In conclusion, the survey to evaluate the "Leaflet for Artificial Intelligence Systems" provided essential insights from a diverse group of AI developers, researchers, and users, ensuring a comprehensive evaluation of the artefact's effectiveness. The results revealed strong support for most of the sections. Feedback from the participants, especially through the qualitative responses, helped to refine the Leaflet to better address the needs of various stakeholders, ensuring it is practical, user-friendly, and aligned with ethical standards. This evaluation process was crucial in enhancing the Leaflet's relevance and applicability for real-world AI governance.

8. LIMITATIONS AND THREATS TO VALIDITY

In this research, an artefact was proposed through the Design Science Research (DSR) process. The following discussion examines the limitations associated with implementing this process, focusing on internal validity, external validity, and reliability.

Concerning **Internal Validity,** which refers to the extent to which the researcher accurately captures reality and reflects it in the study's findings (Merriam, 1995; Merriam, 2009), it is essential to address potential threats that could undermine the causal inferences drawn from the study.

A critical threat to the internal validity of this study lies in the presence of **confounding variables**. These factors, not accounted for in the research design, can influence participants' responses and potentially obscure the true effectiveness of the Leaflet for Artificial Intelligence Systems (LAIS). For instance, participants' prior experiences with AI ethics or familiarity with similar documentation may have shaped their evaluations independently of the LAIS's actual content or design. When such variables are not controlled, they can confound the results, leading to incorrect conclusions about the effectiveness and impact of the artifact.

To mitigate these concerns, the study intentionally incorporated a diverse participant profile, ensuring a broad spectrum of expertise and perspectives. Participants were categorized into three distinct groups to reflect varied backgrounds and professional experiences: **users and researchers not in computing**, **developers in private organizations**, and **academic researchers in computing**. This diversity enriched the data collection process by incorporating insights from different domains, enhancing the depth and **reliability** of the findings.

The inclusion of participants from diverse professional and disciplinary backgrounds ensured that the evaluations of the LAIS captured a wide range of considerations, from technical details to societal implications.

Another threat to validity is **Instrument bias**, which refers to the potential distortion in research results caused by poorly designed or ambiguously interpreted survey instruments. In this study, the primary instrument was a survey used to evaluate the Leaflet for Artificial Intelligence Systems (LAIS). To mitigate the risk of instrument bias, we implemented a multi-faceted approach to ensure the survey design's clarity, reliability, and appropriateness. First, the questions were carefully crafted to avoid ambiguity and leading language that might influence participants' responses.

Furthermore, we included open-ended questions alongside structured Likert-scale items to allow participants to provide nuanced feedback, ensuring that their perspectives were not constrained by predefined response categories. These measures collectively enhanced the validity of the instrument, ensuring that it accurately captured participants' evaluations and minimized the risk of bias impacting the study's findings.

External Validity refers to the extent to which the survey results can be generalized beyond the specific context of the study. When evaluating the survey used to validate the *Leaflet for Artificial Intelligence Systems* (LAIS), several potential threats to external validity should be considered:

A) Sample Representativeness - Survey participants were selected based on their expertise and involvement in artificial intelligence. While appropriate for validating the LAIS, this selection may not fully represent the broader population of AI users and developers. This limitation could affect the generalisability of the findings to contexts with differing demographic characteristics, experience levels, or industry-specific knowledge.

B) Selection Bias -The voluntary nature of survey participation introduces the potential for selection bias. Participants who opted to engage in the study may have had a particular interest in AI ethics or responsible AI practices, potentially skewing the results. This self-selection bias may limit the applicability of the findings to a more diverse or less engaged population.

c) Contextual Differences - Participants may operate within specific cultural, regulatory, and organizational contexts that shape their perceptions of AI ethics and responsibility. These contextual differences could impact the generalisability of the survey results to other regions or industries with distinct legal frameworks, ethical norms, or organizational practices.

To mitigate these threats to external validity, future research should involve a broader and more diverse sample of participants, encompassing individuals from varied cultural, industrial, and regulatory contexts. Additionally, conducting longitudinal studies or replicating the survey in different settings could provide more robust evidence for the generalisability of the LAIS evaluation.

9. CONCLUSION

The transparency principle allows stakeholders, including developers, users, and regulators, to understand how AI systems operate, including their design, decision-making processes, and associated risks. Transparency not only builds trust but also enables the identification and mitigation of errors or biases, fostering more Ethical AI deployment. The absence of a clear decision-making process often leads to negative outputs, as the system operates as a “black box”.

In this context, an artefact, the *Leaflet for Artificial Intelligence Systems* (LAIS), was designed to enhance transparency and provides a practical artefact to bridge the gap between abstract ethical guidelines and their real-world applications, ensuring that AI systems align with societal values and expectations.

The main contribution of this dissertation is the novel artifact called Leaflet for Artificial Intelligence Systems. Also, it enhances understanding of who is responsible when a negative output is generated by an AI system and develops a transparent, objective, and user-friendly document.

The artefact is grounded in the previously studied principle of transparency, which forms the foundation of Ethical AI. Unlike constructs or theoretical models, the LAIS is a practical artefact, based on its layout on the pharmaceutical leaflet, designed for real-world applications. Also, its items and sections were crafted based on the AI Systems legislation and the current state of the art. To evaluate the suitability of the LAIS, we evaluated it utilizing a survey with 18 specialists from different backgrounds. This research distinguishes itself from many others by offering a solution that can be implemented in practice. It addresses the responsibility gap while contributing to both academia and industry by enhancing clarity, transparency, and accountability in AI governance.

Future research may perform longitudinal studies, tracking the application of the artefact within a company, its user acceptance rate, user readership rate, and the number of legal cases following the artefact's implementation. This approach allows researchers to apply the artefact and analyze its impact on transparency.

These future works may be able to precisely investigate the transparency rates between the Leaflet for Artificial Intelligence Systems and related documents, such as Ethical AI guidelines and frameworks.

REFERENCES

2022) Thinking responsibly about responsible AI and ‘the dark side’ of AI, European Journal of ’22: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, ACM pp. 227–236.

A. Badawy, E. Ferrara and K. Lerman, "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 2018, pp. 258-265, doi: 10.1109/ASONAM.2018.8508646.

Acemoglu, Daron, and Pascual Restrepo. 2019. "Automation and New Tasks: How Technology Displaces and Reinstates Labor." Journal of Economic Perspectives, 33 (2): 3-30.DOI: 10.1257/jep.33.2.3

Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who are the Stakeholders? In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22). Association for Computing Machinery, New York, NY, USA, 227–236. https://doi.org/10.1145/3514094.3534187

AI ACT (2023a): Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. [online] [Retrieved December 18, 2023] Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

Alexy, Robert. 1989. A Theory of Legal Argumentation. The Theory of Rational Discourse as Theory of Legal Justification. Oxford: Oxford University Press.

Alexy, Robert. 1998. Law and Correctness. In Current Legal Problems. Ed. M. D. A. Freeman. Oxford: Oxford University Press.

Amershi, Saleema & Inkpen, Kori & Teevan, Jaime & Kikin-Gil, Ruth & Horvitz, Eric & Weld, Dan & Vorvoreanu, Mihaela & Fourney, Adam & Nushi, Besmira & Collisson, Penny & Suh, Jina & Iqbal, Shamsi & Bennett, Paul. (2019). Guidelines for Human-AI Interaction. 1-13. 10.1145/3290605.3300233.

Ampatzoglou, Apostolos & Bibi, Stamatia & Avgeriou, Paris & Chatzigeorgiou, Alexander. (2020). Guidelines for Managing Threats to Validity of Secondary Studies in Software Engineering. 10.1007/978-3-030-32489-6_15.

Aristotle (2004). The Nicomachean ethics. Translated by J. A. K. Thomson. London: Penguin.
Badawy, E. Ferrara and K. Lerman, "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign," *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Barcelona, Spain, 2018, pp. 258-265, doi: 10.1109/ASONAM.2018.8508646.
keywords: {Twitter;Voting;Media;Interference;Tools;Text analysis;Social media manipulation;Russian trolls;Bots;Misinformation},
Bauer, W. A., & Dubljević, V. (2020). AI assistants and the paradox of internal automaticity. *Neuroethics, 13*(3), 303–310. https://doi.org/10.1007/s12152-019-09423-6

Bäuml, F.H. (1980). Varieties and Consequences of Medieval Literacy and Illiteracy. Speculum, 55, 237 - 265.

Bawden, D. and Robinson, L. (2002), Promoting literacy in a digital age: approaches to training for information literacy. Learned Publishing, 15: 297-301. https://doi.org/10.1087/095315102760319279

Benjamins, R. A choices framework for the responsible use of AI. AI Ethics 1, 49–53 (2021). https://doi.org/10.1007/s43681-020-00012-5

Benjamins, V. Richard & Viñuela, Yaiza & Alonso, Chema. (2023). Social and ethical challenges of the metaverse. AI and Ethics. 3. 1-9. 10.1007/s43681-023-00278-5.

Benjamins, V. Richard. (2020). Towards organizational guidelines for the responsible use of AI.

Bentham, Jeremy (1780). An Introduction to the Principles of Morals and Legislation. New York: Dover Publications. Edited by J. H. Burns & H. L. A. Hart.

Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. Proc. ACM Hum.-Comput. Interact. 5, CSCW1, Article 7 (April 2021), 23 pages. https://doi.org/10.1145/3449081

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais. Diário Oficial da União, Brasília, DF, 15 ago. 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 31 de agosto de 2023.

BRASIL. Projeto de Lei do Senado nº 86, de 2021. Projeto de Lei sobre Inteligência Artificial. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/157233. Acesso em: 24 mar. 2024.

Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. Nature Machine Intelligence (2019), 1–7.

Caldwell, M. & Andrews, Jerone & Tanay, Thomas & Griffin, Lewis. (2020). AI-enabled future crime. Crime Science. 9. 10.1186/s40163-020-00123-8.

Caldwell, M.; Andrews, J.T.A.; Tanay, T.; Griffin, L.D. AI-enabled future crime. Crime Sci. 2020, 9, 1–13, https://doi.org/10.1186/s40163-020-00123-8.

Clarke, Roger. (2019). Principles and business processes for responsible AI. Computer Law & Security Review. 35. 10.1016/j.clsr.2019.04.007.

Cortese, João & Cozman, Fabio & de Lucca-Silveira, Marcos & Bechara, Adriano. (2022). Should explicability be a fifth ethical principle in AI ethics?. AI and Ethics. 3. 1-12. 10.1007/s43681-022-00152-w.
da Silva, L., & Seno, E. (2023). Ethics in AI: how software development companies in Brazil deal with the ethical implications of AI technologies. In *Anais do XX Encontro Nacional de Inteligência Artificial e Computacional*, (pp. 156-168). Porto Alegre: SBC. doi:10.5753/eniac.2023.233866

de Laat PB. Companies Committed to Responsible AI: From Principles towards Implementation and Regulation? Philos Technol. 2021;34(4):1135-1193. doi: 10.1007/s13347-021-00474-3. Epub 2021 Oct 6. PMID: 34631392; PMCID: PMC8492454.

Deshpande, Advait and Sharp, Helen (2022). Responsible AI Systems: Who are the Stakeholders? In: AIES

Dignum, V. (2019). Ethical Decision-Making. In: Responsible Artificial Intelligence. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, Cham. https://doi.org/10.1007/978-3-030-30371-6_3

Dignum, Virginia (2019). Responsible Artificial Intelligence: How to Develop and Use Ai in a Responsible Way. Springer Verlag.

DIVINO, S. ESTRATÉGIA BRASILEIRA DE INTELIGÊNCIA ARTIFICIAL (EBIA) E POLÍTICAS PÚBLICAS: PROPOSTAS PARA EFETIVAÇÃO DOS EIXOS LEGISLAÇÃO, REGULAÇÃO E USO ÉTICO E GOVERNANÇA DE IA. E-Legis − Revista Eletrônica do Programa de Pós-Graduação da Câmara dos Deputados, Brasília, DF, Brasil, v. 15, n. 39, p. 45–78, 2022. DOI: 10.51206/elegis.v15i39.797. Disponível em: https://e-legis.camara.leg.br/cefor/index.php/e-legis/article/view/797. Acesso em: 2 abr. 2024.

Domínguez Figaredo, D., & Stoyanovich, J. (2023). Responsible AI literacy: A stakeholder-first approach. Big Data & Society, 10(2). https://doi.org/10.1177/20539517231219958

Dresch, Aline & Lacerda, Daniel & Antunes, Junico. (2015). Design Science Research: Método de Pesquisa para Avanço da Ciência e Tecnologia. 10.13140/2.1.2264.2885.

Duggan, Lawrence. (1989). Was Art Really the “Book of the Illiterate”?. Word & Image. 5. 227-251. 10.1080/02666286.1989.10435406.

Dworkin, Ronald. 1978. Taking Rights Seriously. London: Duckworth.

Dworkin, Ronald. 1986a. Law's Empire. London: Fontana.

Dworkin, Ronald. 1991. On Gaps in the Law. In Controversies about Law's Ontology. Ed. P. Amselek and N. MacCormick. Edinburgh: Edinburgh University Press.

Emirena, Luciana & Almeida, Mauricio. (2019). Design Science: estudo de um campo teórico. Brazilian Journal of Information Science. 13. 0.36311/1981-1640.2019.v13n3.07.p68.

Everett N. Literacy from Late Antiquity to the Early Middle Ages, c. 300–800AD. In: Olson DR, Torrance N, eds. The Cambridge Handbook of Literacy. Cambridge Handbooks in Psychology. Cambridge University Press; 2009:362-385.

Floridi, L. The Fight for Digital Sovereignty: What It Is, and Why It Matters, Especially for the EU. Philos. Technol. 33, 369–378 (2020). https://doi.org/10.1007/s13347-020-00423-6

Floridi, L. Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. Philos. Technol. 32, 185–193 (2019). https://doi.org/10.1007/s13347-019-00354-x

Floridi, Luciano, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, and Christoph Luetge et al. 2018. AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines 28: 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Francés-Gómez, Pedro (2023). Ethical Principles and Governance for AI. In Francisco Lara & Jan Deckers (eds.), Ethics of Artificial Intelligence. Springer Nature Switzerland. pp. 191-217.

Gold, N.E. Virginia Dignum: Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. *Genet Program Evolvable Mach* 22, 137–139 (2021). https://doi.org/10.1007/s10710-020-09394-1

Han, B.-C. (2022). Infocracy (1st ed.). Polity Press. Retrieved from https://www.perlego.com/book/3566687/infocracy-digitization-and-the-crisis-of-democracy-pdf (Original work published 2022)

Hancock PA, Lee JD, Senders JW. Attribution Errors by People and Intelligent Machines. Hum Factors. 2023 Nov;65(7):1293-1305. doi: 10.1177/00187208211036323. Epub 2021 Aug 13. PMID: 34387108.

Hart, Hla (1961). The concept of law. New York: Oxford University Press.

Herbert A. Simon, 1996. "The Sciences of the Artificial, 3rd Edition," MIT Press Books, The MIT Press, edition 1, volume 1, number 0262691914, February.

HERNÁNDEZ SAMPIERI, R.; FERNÁNDEZ COLLADO, C.; BAPTISTA LUCIO, P. Metodologia de pesquisa. 3.ed. São Paulo: McGraw-Hill. 2006.

Hevner, Alan & Parsons, Jeffrey & Brendel, Alfred & Lukyanenko, Roman & Tiefenbeck, Verena & Tremblay, Monica & Brocke, Jan vom. (2024). Transparency in Design Science Research.

Hevner, Alan & R, Alan & March, Salvatore & T, Salvatore & Park, & Park, Jinsoo & Ram, & Sudha,. (2004). Design Science in Information Systems Research. Management Information Systems Quarterly. 28. 75-https://doi.org/10.1038/s42256-019-0088-2

Holmström, Jan & Ketokivi, Mikko & Hameri, Ari-Pekka. (2009). Bridging Practice and Theory: A Design Science Approach. Decision Sciences. 40. 65-87. 10.1111/j.1540-5915.2008.00221.x.

Improved Manipulation Algorithms for District-Based Elections Ramoni O. Lasisi

Jedličková, A. Ethical approaches in designing autonomous and intelligent systems: a comprehensive survey towards responsible development. *AI & Soc* (2024). https://doi.org/10.1007/s00146-024-02040-9
Jedličková, Anetta (forthcoming). Ethical approaches in designing autonomous and intelligent systems: a comprehensive survey towards responsible development. AI and Society:1-14.

Jedličková, Anetta. (2024). Ethical approaches in designing autonomous and intelligent systems: a comprehensive survey towards responsible development. AI & SOCIETY. 1-14. 10.1007/s00146-024-02040-9.

Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. Nat Mach Intell 1, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2

Kant, Immanuel, 1724-1804. (2011). Immanuel Kant : observations on the feeling of the beautiful and sublime and other writings. Cambridge ; New York :Cambridge University Press,

Klatt, Matthias. (2007). Taking Rights less Seriously.A Structural Analysis of Judicial Discretion. Ratio Juris. 20. 506-529. 10.1111/j.1467-9337.2007.00373.x.

L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Turin, Italy, 2018, pp. 80-89, doi: 10.1109/DSAA.2018.00018. keywords: {Artificial intelligence;Computational modeling;Decision trees;Biological neural networks;Taxonomy;Complexity theory;Machine learning theories;Models and systems;Deep learning and deep analytics;Fairness and transparency in data science},

LAKATOS, Eva M. & MARCONI, Marina de A. Fundamentos da Metodologia Científica. São Paulo: Atlas, 1991.

Leikas, J., Koivisto, R., & Gotcheva, N. (2019). Ethical Framework for Designing Autonomous Intelligent Systems. *Journal of Open Innovation: Technology, Market, and Complexity*, *5*(1), 18. https://doi.org/10.3390/joitmc5010018

LEAL, A. L. de C.; ALMENTERO, E.; CUNHA, H.; SOUZA, H. P.; LEITE, J. C. S. do P.. Bula de Software: Uma Estrutura Definida para Promover a Melhoria da Transparência em Software. XV Workshop em Engenharia de Requisitos. 2012.

LIMA, FERNANDO CESAR DE. UMA PROPOSTA DE BULA PARA SOFTWARE. UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ. Dissertação Mestrado, 2015.

Lückmann, Patrick & Feldmann, Carsten. (2017). Success Factors for Business Process Improvement Projects in Small and Medium Sized Enterprises – Empirical Evidence. Procedia Computer Science. 121. 439-445. 10.1016/j.procs.2017.11.059.

Loi, M., & Spielkamp, M. (2021, July). Towards accountability in the use of artificial intelligence for public administrations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 757-766).

Marczyk, Geoffrey R., 1964– Essentials of research design and methodology/Geoffrey Marczyk, David DeMatteo, David Festinger.p. cm.—(Essentials of behavioral science series)Includes bibliographical references and index. ISBN 0-471-47053-8 (pbk.)

Mariana Lenharo, 2024. "AI consciousness: scientists say we urgently need answers," Nature, Nature, vol. 625(7994), pages 226-226, January.

McBride, Neil. (2016). The ethics of driverless cars. ACM SIGCAS Computers and Society. 45. 179-184. 10.1145/2874239.2874265.

McMahon, Sarah & Farmer, Gregory. (2009). The Bystander Approach: Strengths-Based Sexual Assault Prevention With At-Risk Groups. Journal of Human Behavior in the Social Environment. 19. 1042-1065. 10.1080/10911350902990304.

Merriam, S. B. (2009). Qualitative research: A guide to design and implementation. San Francisco, CA: Jossey-Bass.

MI Garage. Ethics Framework - Responsible AI. MI Garage Available at: https://www.migarage.ai/ethics-framework/. (Accessed: 22nd February 2019)

Microsoft. Responsible bots: 10 guidelines for developers of conversational AI. (2018)

Mikalef, Patrick & Boura, Maria & Lekakos, George & Krogstie, John. (2019). Big Data Analytics Capabilities and Innovation: The Mediating Role of Dynamic Capabilities and Moderating Effect of the Environment. British Journal of Management. 30. 272-298. 10.1111/1467-8551.12343.
Miller, Dale E., 2010, *J. S. Mill. Moral, Social and Political Thought*, Cambridge: CUP.

Miller, Dale E., 2011, “Mill, Rule Utilitarianism, and the Incoherence Objection”, in: Eggleston, Ben/Dale E. Miller/David Weinstein (eds.), 2011, *John Stuart Mill and the Art of Life*, Oxford: OUP, 94-116.

Minds and Machines (2019) 29:495–514 https://doi.org/10.1007/s11023-019-09509-3 1 3 A Misdirected Principle with a Catch: Explicability for AI Scott Robbins

Mittelstadt, Brent, Principles Alone Cannot Guarantee Ethical AI (May 20, 2019). Nature Machine Intelligence, November 2019, Available at SSRN: https://ssrn.com/abstract=3391293 or http://dx.doi.org/10.2139/ssrn.3391293

Moser, C., den Hond, F., & Lindebaum, D. (2022). Morality in the age of artificially intelligent algorithms. Academy of Management Learning and Education, 21(1), 139-155. https://doi.org/10.5465/amle.2020.0287

N. Elangovan, E. Sundaravel, Method of preparing a document for survey instrument validation by experts, MethodsX, Volume 8, 2021, 101326, ISSN 2215-0161,

N., Elangovan & Sundaravel, E.. (2021). Method of preparing a document for survey instrument validation by experts. MethodsX. 8. 101326. 10.1016/j.mex.2021.101326.
Nalini, B. (2019). The Hitchhiker’s Guide to AI Ethics. Medium. Retrieved from https://towardsdatascience. com/ethics-of-ai-acomprehensive-primer-1bfd039124b0
Nikolinakos, Nikos. (2023). Ethical Principles for Trustworthy AI. 10.1007/978-3-031-27953-9_3.

OECD (2019), Artificial Intelligence in Society, OECD Publishing, Paris, https://doi.org/10.1787/eedfee77-en.

OECD (2022), "OECD Framework for the Classification of AI systems", OECD Digital Economy Papers, No. 323, OECD Publishing, Paris, https://doi.org/10.1787/cb6d9eca-en.

OECD (2023), "Stocktaking for the development of an AI incident definition", OECD Artificial Intelligence Papers, No. 4, OECD Publishing, Paris, https://doi.org/10.1787/c323ac71-en.

OECD (2024), "Explanatory memorandum on the updated OECD definition of an AI system", OECD Artificial Intelligence Papers, No. 8, OECD Publishing, Paris, https://doi.org/10.1787/623da898-e

OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449

OECD, Recommendation of the Council on Enhancing Access to and Sharing of Data, OECD/LEGAL/0463

OECD, Recommendation of the Council on Principles for Internet Policy Making, OECD/LEGAL/0387

Oosterveld, P., Vorst, H.C.M. & Smits, N. Methods for questionnaire design: a taxonomy linking procedures to test goals. *Qual Life Res* 28, 2501–2512 (2019). https://doi.org/10.1007/s11136-019-02209-6

Oosterveld, Paul & Vorst, Harrie & Smits, Niels. (2019). Methods for questionnaire design: a taxonomy linking procedures to test goals. Quality of Life Research. 28. 1-12. 10.1007/s11136-019-02209-6.

PAPERNOT, Nicolas et al. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security. 2017. p. 506-519.

Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. Hum Factors. 2010 Jun;52(3):381-410. doi: 10.1177/0018720810376055. PMID: 21077562.

Patrick Mikalef, Kieran Conboy, Jenny Eriksson Lundström & Aleš Popovič (2022) Thinking responsibly about responsible AI and ‘the dark side’ of AI, European Journal of Information Systems, 31:3, 257-268, DOI: 10.1080/0960085X.2022.2026621. Stable URL: http://www.jstor.org/stable/2847287. Accessed: 01/08/2013 23:20

Peffers, Ken & Tuunanen, Tuure & Rothenberger, Marcus & Chatterjee, S.. (2007). A design science research methodology for information systems research. Journal of Management Information Systems. 24. 45-77.

PIMENTEL, M.; FILIPPO, D.; SANTOS, T. M. Design science research: pesquisa cient´ıfica atrelada ao design de artefatos. RE@ D-Revista de Educa¸c˜ao a Distˆancia e Elearning, v. 3, n. 1, p. 37–61, 2020. Citado 3 vezes nas p´aginas 18, 50 e 51.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance) (OJ L 119 04.05.2016, p. 1, ELI: http://data.europa.eu/eli/reg/2016/679/oj)

Rittel, H.W.J., Webber, M.M. Dilemmas in a general theory of planning. Policy Sci 4, 155–169 (1973). https://doi.org/10.1007/BF01405730

Robbins, S., & Henschke, A. (2017). The value of transparency: Bulk data and authoritarianism. Surveillance & Society., 15(3/4), 582–589. https://doi.org/10.24908/ss.v15i3/4.6606.

Scott Robbins. 2019. A Misdirected Principle with a Catch: Explicability for AI. Minds Mach. 29, 4 (Dec 2019), 495–514. https://doi.org/10.1007/s11023-019-09509-3

Seifert, Johanna ; Friedrich, Orsolya & Schleidgen, Sebastian (2022). Imitating the Human. New Human–Machine Interactions in Social Robots. NanoEthics 16 (2):181-192.

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. Journal of Database Management, 31(2), pp. 74-87. IGI Global. The definitive version is available at https://doi.org/10.4018/JDM.2020040105

Siau, Keng & Wang, Weiyu. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. Journal of Database Management. 31. 74-87. 10.4018/JDM.2020040105.4

Simon, H. A.(1981)The sciences of the artificial. 3. ed. Cambridge:MIT Press, 1981

Siqueira, Gustavo. (2015). “O parecer de Kelsen sobre a Constituinte brasileira de 1933-1934” / "The Kelsen`s work about Brazilian Constituent 1933-1934". Revista Direito e Práxis. 6. 10.12957/dep.2015.15911.

Sony. Sony Group AI Ethics Guidelines. (2018).

Stepanski, I., & Costa, M. E. (2012). Aspectos comportamentais na gestão de pessoas. Curitiba: IESDE Brasil S.A.

Taylor, I. Is explainable AI responsible AI?. *AI & Soc* (2024). https://doi.org/10.1007/s00146-024-01939-7

The Japanese Society for Artificial Intelligence. The Japanese Society for Artificial Intelligence Ethical Guidelines. (2017).

The Public Voice. Universal Guidelines for Artificial Intelligence. The Public Voice (2018). Available at: https://thepublicvoice.org/ai-universal-guidelines/. (Accessed: 21st February 2019)

Tieto. Tieto’s AI ethics guidelines. (2018).

Tigard, D.W. There Is No Techno-Responsibility Gap. Philos. Technol. 34, 589–607 (2021). https://doi.org/10.1007/s13347-020-00414-7

Unger, R. M. (1983). The Critical Legal Studies Movement. Harvard Law Review, 96(3), 561–675. https://doi.org/10.2307/1341032

UNIÃO EUROPEIA. Regulation of the European Parliament and of the Council Laying Down Harmonized Rules on Artificial Intelligence (Artificial Intelligence Act) [Regulation (EU) 2023/236], 2023. Disponível em: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf. Acesso em: 24 mar. 2024

Van Aken, J. E. (2011). The Research design for Design Science Research in management. Eindhoven: [s.n.], 2011.

Varieties and Consequences of Medieval Literacy and Illiteracy. Franz H. Bäuml. Source: Speculum, Vol. 55, No. 2 (Apr., 1980), pp. 237-265. Published by: Medieval Academy of AmericaVerbal/Visual Enquiry, 5:3, 227-251, DOI: 10.1080/02666286.1989.10435406

Verdiesen, Ilse & Santoni de Sio, Filippo & Dignum, Virginia. (2021). Accountability and Control Over Autonomous Weapon Systems: A Framework for Comprehensive Human Oversight. Minds and Machines. 31. 10.1007/s11023-020-09532-9.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. Science Robotics, 2(6), eaan6080.

Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer. https://doi.org/10.1007/978-3-662-43839-8

Wieringa, Roel. (2009). Design science as nested problem solving. Nuclear Instruments & Methods in Physics Research Section A-accelerators Spectrometers Detectors and Associated Equipment - NUCL INSTRUM METH PHYS RES A. 10.1145/1555619.1555630.

Wieringa, Roel. (2014). Design Science Methodology for Information Systems and Software Engineering. 10.1007/978-3-662-43839-8.

Winfield, Alan & Jirotka, Marina. (2018). Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences. 376. 20180085. 10.1098/rsta.2018.0085.

Zech, Herbert. (2021). Liability for AI: public policy considerations. ERA Forum. 22. 10.1007/s12027-020-00648-0.

Zhou, Jianlong & Chen, Fang & Berry, Adam & Reed, Mike & Zhang, Shujia & Savage, Siobhan. (2020). A Survey on Ethical Principles of AI and Implementations. 3010-3017. 10.1109/SSCI47803.2020.9308437.

APPENDIX 1

LEAFLET FOR ARTIFICIAL INTELLIGENCE SYSTEMS

Preamble

Artificial intelligence (AI) systems are being widely adopted by society in daily activities. Alongside their benefits, numerous issues have emerged from their use. Observing that pharmaceutical leaflets provide patients with clear and precise information about medications, including their use, restrictions, and other relevant factors, we identified a gap in the software industry. While software systems often come with licenses, these documents are generally verbose and lack sufficient informative value. Thus, we conceptualized a "Leaflet for AI Systems," grounded in the principle of transparency. The artifact aims to delineate stakeholder responsibilities by promoting transparency within the system.

To exemplify the potential content of the leaflet, we utilized a selection of AI systems. The information provided in the leaflet about these systems is purely illustrative or fictional, serving only a didactic purpose. The fictícional AI systems included:

- **ROUTES - fictitious route suggestion system for vehicles**
- **IMAGE - fictitious medical imaging diagnostic system**
- **CREDIT - fictitious credit granting analysis system**
- **MUSIC – fictitious automatic music composition system**

Leaflet for Artificial Intelligence Systems

SECTION I: GENERAL INFORMATION

1. Name of the System: Commercial Name of the Artificial Intelligence System. Example: ROUTES
2. Purpose of the System: Description of the impact of the Artificial Intelligence System on society. For example, “ROUTES aims to improve traffic in cities to offer a better quality of life for drivers”.
3. Developed by: Name of the company developing the AI System. Example: ROUTES Inc.

SECTION II: INSTRUCTIONS OF USE
1. System’s Utilization: Description of its main system functionalities and the recommended way to use the AI system. For example, “by sharing geolocated information from different drivers, superimposed on local maps, ROUTES suggests better routes taking traffic intoaccount.t”.
2. Target Audience: Group of people for whom the system was developed. For example, “ROUTES is a system aimed at vehicle drivers.”.
3. Recommendation of Use: Informs the need for human checking of the outputs generated by the AI system. For example, “the ROUTES system does not check whether the suggested routes pass through dangerous areas of the city, therefore, make sure that the route will not bring you unnecessary risks”
4. Level of System Autonomy: Degree of autonomy of the AI system in relation to the need for human intervention in decision-making. For example, “the IMAGE System works in

“human in the loop” mode, which means that its conclusions must imperatively be validated by a medical professional”.

SECTION III. RELIABILITY

1. Error Rate: Report error rates/system accuracy. Examples: “The error rate of the IMAGE system is 95%, measured by the ROC curve and AUC”; “The Template Matching method was used to evaluate the accuracy rate of the IMAGE system, which is 90%”

2. Bias Test: These are tools and methods used to test possible biases, such as gender, age, racial or any other relevant discrimination, reporting the results. For example, “In the development of the CREDIT system, a gender bias test was 54 conducted using the GenBit algorithm, revealing that the model demonstrates an accuracy of 85% for male applicants and 75% for female applicants”

SECTION IV: DATA LIFECYCLE

1. Data Collection: Explains what data the AI system collects, in addition to explicitly requiring user acceptance. Example: “All text that the user types when interacting with ChatGPT is stored by the system and can become public.”

2. Data Storage: Indicates how data is stored, for how long, the level of security, and whether it is anonymized. Example: “The CREDIT system follows ISO 27001 cybersecurity protocols when storing data. It anonymizes data and deletes sensitive data; among other LGPD requirements and ANPD guidelines. Data is stored for an unlimited time but can be deleted at the owner’s request.”

3. Scope of Use: Provides details about the types of data that will be collected (texts, images, audio, etc.), as well as the purpose of this collection. Example: “The ROUTES system collects geolocation information from drivers in real-time while using the application. Such data is used to improve the ROUTES system”.

4. Data Sharing: List with whom the collected data can be shared, and whether it can be used by other systems within the same company. Example: “The images and diagnoses collected by the IMAGE system, in addition to serving to improve the system, can be used by other systems from the same developer. However, such data will be passed on to any third party under any circumstances.”

5. Data Disposal: Informs how the user can request the modification or disposal of their data. Furthermore, it also informs the minimum time needed for the AI system to relearn without using the deleted or modified data. Example: “The company CREDIT Inc adopts data disposal practices that ensure the safe and permanent removal of user information when requested. Deletions and modifications of personal data can be made via the website www.credit.com/sac or by calling SAC at 0800 878 787. At most every 3 months, Credit's machine learning models are retrained to take into account changes in their databases.”

6. Intellectual Property: Explains that it has obtained authorization from the owners of the data used in training the AI system and indicates a path for complaints. Example: “All musical examples used to train the MUSIC system were either in the public domain or were duly authorized for this purpose by whoever owned their intellectual property. For any complications, please contact www.music.com/sac”.

SECTION V: RESPONSIBILITIES

1. Forum and Conflict Resolution: Forum and Channel for conflict resolution. Example: The forum for resolving conflicts involving the CREDIT system will be in the City of Recife, Brazil. Extrajudicial notifications must be sent to juridico@credit.com”

2. Developer and Operator Responsibility (statement): “Our commitment as developers and operators of the AI system is to ensure ethical, transparent, and 58 responsible treatment of user data. In this way, our objective is to generate results that benefit society, minimizing possible negative impacts generated by the system, as well as seeking to mitigate any possible form of discrimination generated by AI.”

12

APPENDIX 2

EVALUATION FORM

EVALUATION FOR ARTIFICIAL INTELLIGENCE SYSTEMS

Hello! Firstly, we thank you for your participation in our research. This is the evaluation form for a proposal for a Leaflet for AI Systems developed as part of Maria Renata Gois' dissertation for the postgraduate program at the UFPE Informatics Center. The objective of which is to make understanding more transparent regarding the responsibility for negative outputs generated by AI systems. To evaluate our artefact we need you to respond to our evaluation, the estimated time is 20 minutes. Thank you for your collaboration.

1.Do you qualify as:

( )Person who develops artificial intelligence in a company
( ) Researcher in Artificial Intelligence
( ) Exclusively User of Artificial Intelligence Systems

2. If you are in the Artificial Intelligence Development area, you have

( )Up to 2 years of experience
( )2-5 years of experience
( )5-10 Years of Experience
( )+10 years of experience

3. If you are a User, your area of professional activity

____

General Opinion Section on the Leaflet
Assessment in this session is based on the Likert scale.
Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

Rate each proposition from 1-5 regarding the sections of the Bulletin for AI Systems

4. The Insert must contain section 2.1 Use of the System that describes the main functionalities of the system.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

5. The Insert must contain section 2.2 Target Audience that informs which group of people the system was developed for

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

6.The Insert must contain section 2.3 Recommendations for Use, which informs the user's need to pay attention to the target audience, time of use and the need for human checking of the outputs generated by the system.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree
7.The Insert must contain section 2.4 Classification of the Level of Autonomy that informs the degree of autonomy of the AI system in relation to the need for human intervention in decision-making

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

8.The Insert must contain section 3.1 Error Rate that reports the error and accuracy rates of the system.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

9.The Insert must contain section 3.2 Bias testing that contains which biases were tested

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

10.The Insert must contain section 3.3 Potential Risks which contains a descriptive list of potential risks arising from the use of the system.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

11.The Insert must contain section 4.1 Data Collection that defines what data is, lists what data the AI system collects, and explicitly requires user acceptance.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

12.The Insert must contain section 4.2 Data Storage which explains how to store it, for how long, and for what purpose the data should be used.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree

Rating 5: Totally Agree

13. The Leaflet must contain section 4.3 Scope of Use of Data which lists how the data may be used.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

14.The Insert must contain section 4.4 Data Sharing which lists with whom the collected data may be shared, who the system's data operators are and whether this data may be used by other systems of the same company.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

15. The Insert must contain section 4.5 Data Disposal which informs how the user can request the modification, alteration or disposal of their data. *

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

16. The Insert must contain section 4.6 Intellectual Property that informs channels that the user can contact if they notice a violation of intellectual property in the database used by the AI system.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

17. The Bulletin must contain section 5.1 Forum and Conflict Resolution which indicates how and where conflicts will be legally resolved.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

18. The Insert must contain section 5.2 Developer Responsibility which declares its commitment to being ethically and legally aligned, seeking transparency regarding the understanding of the AI system.

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree
19. It is pertinent that there is a Leaflet for AI Systems *

Note 1: Totally Disagree

Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

20. The Information Leaflet for AI Systems is innovative

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

21. The Leaflet for AI systems as it is ethically aligned

Note 1: Totally Disagree
Note 2: Partially disagree
Note 3: Neither Agree nor Disagree
Note 4: Partially Agree
Rating 5: Totally Agree

22. What do you think of this Initiative to create the Leaflet for AI Systems?
____

23.Would you add any items to the Information Leaflet for AI Systems?

____

24. Would you remove anything from the Information Leaflet for AI Systems? *

____

25.Would you change anything in the Information Leaflet for AI Systems?
____

26. If you want to comment anything else, enter it here

____

APPENDIX 3

Ethical Principle	AI Act
Transparency	Art 13, (a) AI Act- Identify the contact details of the provider Art 3, AI Act- Provider of AI System Propose 6 AI ACT- The notion of AI system should be clearly defined Art 3, AI Act- Provider of AI System ART 7, (b), AI Act- the extent to which an AI system has been used or is likely to be used Art 13, B, IV AI Act- groups of persons on which the system is intended to be used Propose 11, AI Act- AI system should fall within the scope of this regulation
Explicability	Propose 6 AI ACT- The notion of AI system should be clearly defined Art 13 (b) AI ACT I- Its purpose
Explicability, Transparency	ART 7, (b), AI Act- the extent to which an Ai system has been used or is likely to be used Propose 43 AI Act- requirements should be applied regards data sets
Explicability Transparency, Autonomy, Beneficence	ART 7, (G), AI act- the extent of the output produced by AI system is reversible, or have an impact
Transparency, Explicability, non-maleficence	Propose AI Act- Technical inaccuracies Propose AI Act- Robustness is a key requirement Propose 51 AI Act- Cybersecurity Propose AI Act- Ensure that providers take in account AI systems risksPropose14 AI Act- Scopus of the risks that AI system can generate Propose 28 AI Act- Adverse outcomes Propose 72 AI Act- Sandboxes and controlled experiments Propose 78 AI Act- Ensure that providers take into account AI systems risks Art 10, (f) AI Act- Examination in view of possible biases Propose 14 AI Act- Scopus of the risks that AI systems can generate
Transparency, explicability, beneficence, non maleficence, justice	Propose 43 AI Act- requirements should be applied regards data sets Propose 46- Information about data lifecycle and compliance Propose 78 AI Act- Ensure that providers take into account AI systems risks

APPENDIX 4

GDPR	Ethical Principle
Art 1- any natural person Art 27- Non applied to dead Art 32- Agreement	Explicability, Transparency
Art 1, 1- any natural person Art 32- Agreement	Explicability Transparency, Autonomy, Beneficence
Art 26- data anonymity Art 47, 50, 44- data sharing art 23- data finality, Scopus, and storage	Transparency
Art 39- scopus of utilisation	Transparency, non-maleficence
Art 51- operator responsibilities Art 82- administrative sanctions art 24- controller responsibility art 40, 2, k- judicial process	Justice,autonomy, Explicability, Beneficence.

APPENDIX 5

LGPD	Ethical Principle
Art 1- any natural person art 9, I- data treatment	Transparency
art 33- data international sharing art 15- data treatment	Transparency, non-maleficence, autonomy
Art 37,38,39,40- obligations art 50- governance and good practices art 52, 53- administrative sanctions Art 55- ANPD Conflict Resolution	Justice,autonomy,Explicability, Beneficence, transparency
Art 37,38,39,40- obligations of the operator and the controller	Transparency, justice, non-maleficence

APPENDIX 6

Ethical Principle	PL of AI
Explicability	Art 7º, IV PL de IA - Finality of the System
Transparency	Art 5 PL de IA - Informations about system interactions Art 7º, IV PL de IA - Finality Information of the System
Explicability, Transparency	Art 5, I PL de IA- Information about system interactions
Explicability Transparency,Autonomy, Beneficence	Art 7º, II PL de IA- Recommendations and consequences of use Art 7º, V PL de IA – Data categorization Art 9, I PL de IA- Data finality
Explicability, Transparency, Beneficence	Art. 3º, III, PL de IA Art 8º, II PL de IA- Level of System Autonomy Art 8º, V, PL de IA- Possibility of Human Interference EMENDA ART 9, I E II PL DE IA- System autonomy and its risks Art 13, PL de IA- System risks Art 3, VII PL de IA- Systems Trust
Transparency, Explicability, non-maleficence.	Art 5º, V PL de IA - non-discrimination, prejudices Art 7 VI, PL de IA- Security measures
Transparency,justice, explicability	Art 5º, VI PL de IA - Private and Data protections Art 13 PL de IA - Level of Risk Art 3º, VIII PL de IA – due process of law Art 5º, III, PL de IA – decisions
Transparency, non-maleficence	Art 21, III, PL de IA - data protocols
Justice, Autonomy, transparency, non-maleficence	Art 3º, X PL de IA- responsibilities Art 9 PL de IA - data modifications and decisions Art 27 PL de IA – civil liability

APPENDIX 7

SECTION	ETHICAL PRINCIPLE	DEFINITION
1. General Information	Transparency	It will provide general information about the AI System.
1.1 Name of the System	Transparency	It provides the system name
1.2 Puporse of the System	Explicability	Defines the purpose of the system, what it intends to do
1.3 Developed by	Transparency	Provides the name of the company that developed the system
2. Indications of Use	Transparency	Refers to the mode of use and the target audience.
2.1.Use of the System	Explicability, Transparency	Description of the system's main functionalities and recommendations for its use.
2.2 Target Audience	Transparency	Group of people for whom the system was developed
2.3 Recommendations of Use	Explicability Transparency,Autonomy, Beneficence	Informs the need for human verification of the outputs generated by the AI system.
2.4 Level of System Autonomy	Explicability, Transparency, Beneficence	Degree of autonomy of the AI system.
3.**Trust**	Beneficence, Transparency	Explains the risks involved when using the AI System
3.1 Tax of Errors	Transparency, Explicability, non-maleficence	Report system accuracy/error rates.
3.2 Bias Tests	Transparency, Explicability, non-maleficence.	These are tools and methods used to test possible prejudices, such as gender, age, race or any other relevant discrimination, reporting the results.
4.Data Lifecycle	Transparency,justice, explicability	Set of steps through which data passes from its collection to its eventual disposal.
4.1 Data Collection	Transparency, explicability,, beneficence.	.Explains what data is, what data the AI system collects, and explicitly requires user acceptance
4.2 Data Storage	Transparency, explicability	Explains how data is stored, for how long, level of security and anonymity.
4.3 Scopus of Utilisation	Transparency	Provides details about the types of data that will be collected, as well as the purpose of that collection.
4.4 Data Sharing	Transparency, non-maleficence	List with whom the collected data can be shared and whether it can be used by other systems within the same company.
4.5 Data Removal	Transparency, autonomy, non-maleficence	It informs how the user can request the modification or deletion of their data, in addition, it also informs how long it takes for the AI system to relearn without using that user's data.
4.6 Intellectual Property	Transparency, justice	Indicates channels for reporting intellectual property violations in the AI system database.
5. Responsibilities	Justice, autonomy, transparency	Information that clarifies the obligations, limitations and responsibilities of the negative results of the artificial intelligence system.
5.1 Forum Conflict Resolution	Justice, autonomy,Explicability, Beneficiency.	Information on how users can contact the developer and where to handle legal issues created by the AI system
5.2 Responsibilities of the Operator and the Developer	Justice, Autonomy, transparency, non-maleficence	Statement