AIPA 1/2024

Arbeitspapiere zur Internationalen Politik
und Außenpolitik

Jerome Harrison

International Norm Dynamics of AI Ethics: The Role of the European Union

Lehrstuhl für Internationale Politik und Außenpolitik
Universität zu Köln
ISSN 1611-0072

AIPA 1/2024

Arbeitspapiere zur Internationalen Politik
und Außenpolitik

Jerome Harrison

International Norm Dynamics of AI Ethics: The Role of the European Union

ISSN 1611-0072

Lehrstuhl für Internationale Politik und Außenpolitik
Universität zu Köln, Aachener Straße 75, 50931 Köln

Redaktionelle Bearbeitung: Melanie Seilnacht

Köln 2024

Abstract

This paper examines the influence of the European Union's (EU) approach to Artificial Intelligence (AI) Ethics on the United States (US). The study is based on Constructivism and the concept of Normative Power Europe, which suggests that the EU plays a significant role in shaping international norms, beliefs, and values related to AI Ethics. The research employs a comparative analysis of the EU's and US's approaches to AI Ethics, focusing on the period from 2018 to 2023 through a systematic document analysis. The findings reveal a substantial alignment of the US's approach with that of the EU's particularly regarding the concept of 'trustworthy AI'. The analysis identifies emulation and competition as the primary diffusion mechanisms driving the convergence of ethical perspectives between the EU and the US. Despite this alignment, the US has yet to adopt a legally binding regulatory framework for AI Ethics, relying instead on guidelines and principles. The research highlights the importance of understanding international dynamics in AI Ethics, emphasizing the ongoing need for scrutiny and exploration as the field evolves. The adoption of the EU AI Act and its potential impact highlight the significance of ongoing research and analysis in this field.

Keywords: AI Ethics, AI Governance, European Union (EU), United States (US), International Norm Dynamics, Diffusion Mechanisms

Jerome Harrison studierte Politikwissenschaft (M.A) mit fachlichem Schwerpunkt auf europäischer und internationaler Politik an der Universität zu Köln. Beruflich bewegt er sich durch Stationen bei der Gesellschaft für Internationale Zusammenarbeit und der Fraunhofer-Gesellschaft an der Schnittstelle von Innovation, Gesellschaft und Entwicklungszusammenarbeit.

Kontakt: jerome.harr.son@gmail.com

Table of Contents

List of Tables

List of Abbrevations

AGI	Artificial General Intelligence
AI	Artificial Intelligence
AI HLEG	High-Level Expert Group on Artificial Intelligence
AI RMF	Artificial Intelligence Risk Management Framework
CPO	Causal-Process Observations
EC	European Commission
EU	European Union
EO	Executive Order
GDPR	General Data Protection Regulation
GPAI	Gloabal Partnership for Artificial Intelligence
NGO	Non-Governmental Organisation
NIST	National Institute of Standards and Technology
NPE	Normative Power Europe
OECD	Organisation for Economic Co-operation and Development
OMB	Office of Management and Budget
OSTP	Office of Science and Technology Policy
TTC	Trade and Technology Council
US	United States of America
UK	United Kingdom

International Norm Dynamics of AI Ethics: The Role of the European Union

1 Introduction

In November 2022, the non-profit organization OpenAI, based in the United States, shocked the world by releasing its large language model-based chatbot ChatGPT to the public. ChatGPT is capable of producing large amounts of human-like texts within seconds. It has been widely regarded as 'one of the best artificial intelligence (AI) chatbots ever released to the public' (Roose 2022). ChatGPT is a notable example of the ongoing 'AI boom' (The Economist 2023), which has sparked new ethical discussions about automated systems in public and in research.

However, ethical discussions surrounding AI are not a new phenomenon. In 1942, Isaac Asimov presented 'Three Laws of Robotics' in his science fiction short story 'Runaround', which later appeared in his famous collection 'I, Robot'. These laws formulate ethical provisions regarding robots. In 1950, Alan Turing proposed the 'imitation game' as a test for a machine's ability to exhibit intelligent behavior indistinguishable from that of a human (Turing 1950). This test, commonly referred to as the 'Turing Test', remains relevant today. Five years later, the term 'AI' was coined for the first time in a research proposal. The proposal states that "[t]he study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it." (McCarthy et al. 1955).

Since its inception, the field of AI has undergone several phases of lower and higher public interest, commonly referred to as 'winters' and 'summers'. These phases have been accompanied by corresponding interest in ethical considerations around AI systems (Floridi 2021b). Currently, due to the ongoing AI boom, public interest in AI is extremely high. In 2014, Elon Musk, the founder of Tesla and co-founder of OpenAI, said about AI: "If I had to guess at what our biggest existential threat is, it's probably that" (Gibbs 2014). Physicist Stephen Hawking warned that AI could be "either the best, or the worst thing, ever to happen to humanity" (Hern 2016). Researchers have also highlighted the risks and challenges associated with highly developed AI systems in various aspects of our lives (Brundage et al. 2018) To prevent AI developments from spiraling out of control, the Future of Life Institute published an open letter in March 2023. The letter called for a six-month pause in AI research and development, and was endorsed by public figures such as historian Yuval Noah Harari, Elon Musk, computer scientist Stuart Russel, and Apple co-founder Steve Wozniak (Future of Life Institute 2023; Metz and Schmidt 2023). During a US Senate hearing in May 2023, the CEO of OpenAI stated that "if this technology goes wrong, it can go quite wrong" and called for new regulations and a set of safety standards for AI models (Kang 2023).

Countries and organizations have been addressing the risks and ethical considerations related to AI and have published corresponding ethical principles and guidelines (Jobin et al. 2019; Daly et al. 2019; Schmitt 2022). The European Union (EU) has been a major focus in this regard, particularly after the European Commission (EC) released the first AI-specific legislative draft in April 2021, known as the 'AI Act' (European Commission 2021). While the draft is currently under debate within the EU and is expected to be adopted by the end of 2023 (Sharp 2023). Research is being conducted to discuss its potential impact on the international system (Greenleaf 2021; Birchfield et al. 2022; Feldstein 2023).

Instead of projecting a potential future impact, this thesis aims to evaluate the normative influence that the EU has had in the field of AI Ethics so far. Since 2018, the EU has consistently increased its efforts to deal with AI in an ethically sound manner. The underlying assumption of this thesis is based on the constructivist assumption that the international system is shaped by socially constructed norms, beliefs and values. This thesis is based on the concept of 'Normative Power Europe' (Manners 2002), which suggests that the EU has a unique international normative role and aims to influence and shape international norms, beliefs, and values. The thesis assumes that this is also the case with regard to AI Ethics. In order to evaluate the EU's normative influence on AI Ethics this thesis will specifically compare the EU's approach to AI Ethics to the approach of the US, the current international powerhouse in the field of AI (Tortoise 2023). The central research question of the thesis therefore is: *What is the influence of the EU's AI Ethics approach to AI Ethics?* Hence, this research is comparative in nature and situated within an X-centered research design. The goal is to determine the effect the explanatory variable (the EU's approach to AI Ethics) has on the dependent variable (the US's approach to AI Ethics). In this context, the research question is divided into two parts. First, based on assumptions derived from Constructivism and Normative Power Europe, the thesis aims at identifying *if* the EU's AI Ethics approach has had an influence on the US. For this purpose, this thesis will present an overview of both the EU's and the US's approach to AI Ethics over time, based on a systematic analysis of official documents. These approaches will be compared to identify potential convergences. In case a shift of the US's approach to AI Ethics towards the EU's approach is observed, the second part will seek to identify *why* this shift has occurred. This part is rooted in diffusion theory and, via the method of process tracing, aims to identify potential causal diffusion mechanisms between the EU and the US regarding AI Ethics.

This research holds scientific and societal significance as it provides a comprehensive comparison of the EU and US approaches to AI Ethics over time.

The empirical value of this study can be used for future research, while also testing the validity of theories like Normative Power Europe and diffusion in the highly dynamic and relevant field of AI Ethics. By taking a retrospective view and comparing the EU and the US this thesis addresses an existing gap in academia. Additionally, due to the transformative nature of AI and the ethical risks and challenges involved, this thesis is of societal value, too, as it enhances the understanding of international AI governance and corresponding motivations and relationships.

This thesis will begin by discussing the various facets of AI Ethics and providing an overview of how the topic has been dealt with in research – including an overview of existing research on EU and AI Ethics. It will commence by introducing the guiding theoretical concepts of Constructivism, Normative Power Europe and Diffusion and discuss how, combined, they serve as a valuable theoretical framework for this thesis's research objectives. Subsequently, based on the theoretical framework, testable hypotheses will be developed to guide the empirical analysis. This is followed by the introduction of the method that is applied during the analysis – process tracing – and a systematic operationalization of the hypotheses based on process tracing and document analysis. This operationalization serves as the basis for the subsequent empirical analysis. The following discussion and conclusion will summarize the findings and discuss their implications.

2 AI Ethics

The following scenario is not unlikely in the very near future: An autonomous car is involved in a lethal accident. This scenario raises a question as old as humanity: Who is responsible? Is it the owner? The AI manufacturer? Or perhaps the car itself?

Such moral questions are a at the core of philosophy and have been debated for centuries. Ethics is a discipline within philosophy that aims to answer these questions by identifying what is right and what is wrong. In the process, ethics seeks to observe and understand the underlying motivation and values of human behavior (Boddington 2017; Daly et al. 2019). Within ethics, applied ethics has a practical approach to these moral questions. And within applied ethics, AI Ethics tries to examine and understand the interplay between AI related advancements and their impact on society (Daly et al. 2019).

Two fields have emerged in AI Ethics: Machine ethics and robot ethics. Machine ethics is discussed less commonly than robot ethics – as of now. It might gain importance in the future as it concerns the ethical behavior of AI systems (Winfield et al. 2019; Bostrom and Yudkowsky 2018). Robot ethics, also known as AI Ethics, is about how humans develop, deploy, or use AI systems ethically. Its aim is not only to prevent harm but also to maximize the benefits of AI in an ethical manner. To achieve both goals, many corresponding ethical principles, standards, and best practices have been developed to date (Winfield et al. 2019). Highlighting this dual nature of AI Ethics, Floridi states that AI is concerned with principles of beneficence (ensuring the benefits of AI), non-maleficence (preventing harm), autonomy (respecting human agency and decision-making), justice (ensuring fairness and equity), and explicability (making AI systems transparent and understandable). These principles are necessary for an AI system to be both technologically advanced and ethical (Floridi 2021c). This dual focus on benefit-maximization and harm-prevention shows that the common perception that ethics stifles innovation and research is limited (Boddington 2017).

Dignum (2018) differentiates between three ethical design approaches. Firstly, ethics by design, which requires the AI system to include 'ethical reasoning capabilities' (Dignum 2018). Secondly, ethics in design, referring to the necessity to evaluate the ethical implications of regulatory and engineering methods in the

implementation of AI. And lastly, ethics for design, which requires responsible developers to adhere to ethical standards, procedures, and conduct (Dignum 2018).

Because this thesis is concerned with the role of the EU in international norm dynamics of AI Ethics, it will borrow the definition of AI Ethics from the 'Ethics Guidelines for Trustworthy AI', created by the High-Level Expert Group on Artificial Intelligence (AI HLEG) for the European Commission:

> AI Ethics is a sub-field of applied ethics, focusing on the ethical issues raised by the development, deployment and use of AI. Its central concern is to identify how AI can advance or raise concerns to the good life of individuals, whether in terms of quality of life, or human autonomy and freedom necessary for a democratic society (AI HLEG 2019a).

This definition acknowledges that AI Ethics is about more than just preventing harm; it is also about promoting good. It aligns with Leslie's (2019) definition of AI Ethics as 'a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies' (Leslie 2019). This definition places greater emphasis on the moral implications of AI Ethics.

2.1 Challenges and risks of AI

Today, AI is already demonstrating why it is often referred to as the 'ultimate enabler' (Horowitz 2018) and 'transformative' (Boddington 2017), as it improves our daily lives across multiple domains (Dignum 2018). Nonetheless, a growing number of researchers are dedicated to pointing out risks involved in the development and application of AI, as well as the need for ethical development and application of AI (Bostrom and Yudkowsky 2018; Leslie 2019; Mittelstadt et al. 2016; Dafoe 2018)

In a joint report on the potential malicious use of AI, various AI researchers outlined risks involved with AI. The report considers three AI-related security domains in which risk can occur. The first, digital security, encompasses

cyberattacks, human-, software-, and AI system vulnerabilities. The second, physical security, includes automated drones and other physical systems, cyber-physical systems, and physical systems. The third, political security, is concerned with surveillance, persuasion, and deception (Brundage et al. 2018).

Horowitz (2018) points to the complexities surrounding AI in the international system. Due to the rapid advancements of AI systems, they may have completely reshaped international competition and the existing balance of power, potentially even resulting in an AI arms race. This race could incentivize international actors to neglect risks associated with AI (Horowitz 2018). Other researchers have also identified inherent risks of AI systems. One of the main ethical challenges in the field of AI is the issue of bias and discrimination. AI systems are data driven, and as a result, they may reproduce these issues and thereby intensify social prejudice (Leslie 2019). This is because humans program AI systems and feed their own values and biases into these systems, causing the algorithms to reflect existing societal structures (Mittelstadt et al. 2016). This is especially problematic when AI is perceived as objective (Bostrom and Yudkowsky 2018).

Additionally, the lack of transparency in AI systems presents a major challenge. Due to their complexity, it can be difficult to understand and trace the decision-making process, leading to the common reference of 'black boxes' (Bostrom and Yudkowsky 2018). However, transparency is essential for assigning responsibility in cases of harm or error, as well as for establishing trust in AI systems (Mittelstadt et al. 2016). Robustness and security are further concerns when it comes to AI – particularly in safety-critical applications and sectors. In such cases, both the AI systems and the humans involved are highly vulnerable (Bostrom and Yudkowsky 2018). Furthermore, AI raises privacy concerns. To safeguard the privacy rights of individuals, it is essential to ensure human oversight in the design, development, and deployment of AI systems (Leslie 2019; Floridi et al. 2018). Another challenge mentioned is of a social nature. To maintain strong social

connections and preserve individual experiences, it is important to avoid over-reliance on AI systems (Leslie 2019).

One potential challenge, although not yet present, is the emergence of Artificial General Intelligence (AGI) and super-intelligence. Unlike 'regular' AI, AGI possesses the ability to think and learn beyond pre-programmed rules, which raises concerns about the possibility of superintelligence. This AI has the potential to continuously improve itself without limit. While it could bring huge benefits, it also comes with serious risks (Bostrom and Yudkowsky 2018).

2.2 Governance of AI Ethics

This chapter presents an overview of international approaches to AI governance and the respective research conducted. Butcher and Beridze (2019) define AI governance as 'a variety of tools, solutions, and levers that influence AI applications' (Butcher and Beridze 2019). Floridi (2018) provides a broader definition of digital governance. He defines it as 'the practice of establishing and implementing policies, procedures, and standards for the proper development, use, and management of the infosphere' (Floridi 2018). Winfield and Jirotka, on the other hand, define ethical governance as a 'set of processes, procedures, cultures, and values designed to ensure the highest standards of behavior' (Winfield and Jirotka 2018). While every definition has a different focus, one thing is clear: AI governance is about practical tools and processes to cope with AI systems.

Ethical principles and guidelines related to AI have been the focus of much academic attention, but they are almost always non-binding. Among others, Jobin et al. (2019) and Fjield (2020) conducted analyses of multiple international documents containing ethical principles or guidelines related to AI. Jobin et al. (2019) analyzed 84 documents and found a significant increase in global interest in AI Ethics since 2016. The study identified 11 ethical values and principles. Five of

these 11 values and principles were cited in more than half of the analyzed documents: transparency, justice and fairness, non-maleficence, responsibility, and privacy (Jobin et al. 2019). Similarly, in their analysis of 36 international documents, Fjeld et al. (2020) identified eight key principles: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values (Fjeld et al. 2020).

But where do these principles originate from? Although there is some level of agreement, it is uncertain whether various groups influence each other, develop these principles based on the same research documents, or develop them individually (Whittlestone et al. 2019). Nevertheless, principles are highly valuable in applied ethics as they translate complex ethical issues into elements that can be understood by everyone. This enables a shared commitment to the same values across different countries and sectors. This is important in order to develop ethical standards, international agreements, and regulations based on the principles and values (Whittlestone et al. 2019). Additionally, Papyshev and Yarime (2023) identified three different focal points in their analysis of 31 national AI strategies. They found that post-Soviet bloc European countries, China, and East Asia tend to emphasize 'Development'. EU countries prioritize ‘Control’, while the UK, US and Ireland focus on ‘Promotion’ to encourage innovation (Papyshev and Yarime 2023). According to Schiff et al.`s (2021) findings, private sector recommendations tend to concentrate on client-related and technical concerns, while public and NGO materials tend to be more interactive and engaged with legislative aspects (Schiff et al. 2021). It is noteworthy that in some countries the economic development of AI is prioritized over its societal impact. Furthermore, ethical standards for AI consist of four main components: economic, research, societal, and regulatory considerations (Schmitt 2022). Djeffal et al. (2022) analyzed AI policies from 22 countries and the EU. The authors found that different countries favor different approaches to AI governance: some prefer self-regulation and market-based approaches, while others

combine entrepreneurial and regulatory methods. However, the authors note that regardless of the governance approach, there is a consistent focus on public responsibility. Most governments have shown a willingness to promote AI technologies while also regulating associated risks (Djeffal et al. 2022). In this context, Smuha (2021) suggests that enabling regulations could provide economic incentives, while protective regulations could enforce transparency and ethical standards. This could result in a race to regulate AI (Smuha 2021). According to Garcia (2022), countries will need to cooperate on AI governance instead of individually racing to regulate AI. International organizations, such as the United Nations, will play a crucial role in this matter. Recently, there has been an increase in AI-related collaboration at the OECD and the Global Partnership for AI (GPAI) (Garcia 2022).

2.3 Challenges and risks in AI Ethics

The goal of AI governance is to effectively navigate the field of AI Ethics, which is a complex task with numerous challenges and risks. In this context, researchers have raised several issues and concerns.

Although ethical principles are crucial for AI governance, they are also problematic in some ways. By themselves, they are insufficient to address the challenges and risks associated with AI systems, as outlined in chapter 2.1 of this thesis. Action based on objective principles is necessary. Furthermore, it is important to consistently question and reiterate principles (Rességuier and Rodrigues 2020). However, the lack of clarity in principles can lead to misunderstandings as they may be interpreted differently by different actors. Additionally, the simplicity of the principles can make it challenging to take action based on them (Whittlestone et al. 2019). Mittelstadt (2019) draws a comparison between the challenges faced by AI and medicine. Although medicine aims to

improve health, the field of AI is more complex and often lacks effective methods to translate ethical principles into practice. Mittelstadt (2019) highlights that one of the most significant concerns is the lack of enforceable ethical frameworks in AI, unlike in medicine.

Another issue in AI ethics is bias. Ethical considerations are often based on judgments, which can be biased. Hagendorff (2023) identifies 'role-model risks' and 'bounded ethicality risks' in this context. It is crucial for AI ethicists to actively address their biases and strive to avoid personal beliefs from interfering with their AI-related work (Hagendorff 2023). Hagendorff also highlights a gap between the ongoing ethical debate and the actual technical understanding. He warns against 'non-expert risks´ (Hagendorff 2023), which refers to the danger of becoming too abstract without the necessary technical insight. To avoid this, he suggests that AI ethicists work interdisciplinary (Hagendorff 2023). Additionally, he argues that fairness and privacy are not receiving enough attention in the field (Hagendorff 2022).

The challenge of AI governance and AI Ethics is to be aware of these risks and to mitigate them. This requires extensive ethics training (Hagendorff 2023), transparency (Floridi 2021c), and a culture of continuous questioning and renewal (Rességuier and Rodrigues 2020) .

2.4 The EU's approach to AI Ethics and its international role

This research examines the EU's role in shaping international norms around AI Ethics. The EU's value-driven approach to AI Ethics has received a lot of attention from researchers. This chapter provides an overview of the EU's existing approach to AI Ethics and its international role, identifying gaps that this research aims to address.

In his analysis, Larsson (2020) examines the use of ethics guidelines in AI governance and thereby closely scrutinizes the AI HLEG's Ethics Guidelines for Trustworthy AI for the EC in this context. The author emphasizes the need to transition from principles to processes in AI governance and notes the temporal discrepancy between rapid technological advancements of AI systems and respective legal reforms (Larsson 2020).

Birchfield et al. (2022) explore whether the EU has the potential to become a global leader in ethical and secure AI. The authors emphasize the importance of defining AI for the EU, compare the EU's technical competitiveness to that of the US and China, and draw on the concept of the 'Brussels Effect', which refers to the EU's regulatory influence on the international system. In order to access the European market, companies must comply with European standards and subsequently implement them globally (Bradford 2012). Birchfield et al. highlight the EU's history of influencing non-European entities through standards and legislation, as demonstrated by the General Data Protection Regulation (GDPR) and the potential for the EU's AI Act to exert regulatory influence (Birchfield et al. 2022).

Similarly, Greenleaf (2021) argues that the AI Act could serve as a cornerstone in privacy protection, not only within Europe but also for international companies. He emphasizes the AI Act's potential role (and responsibility) in providing clarity in the otherwise fragmented AI regulatory landscape. Once enacted, the AI Act is likely to influence AI legislation in countries beyond Europe (Greenleaf 2021).

The EU's attempts to influence international AI Ethics, particularly through the AI Act, are examined by Feldstein (2023). He argues that the EU's swift action to pass the AI Act reflects its intention to shape global AI governance. Feldstein acknowledges the complexity of AI norm-building, including issues such as pacing and fragmentation, but recognizes the high potential of the AI Act to establish international norms. However, he raises concerns about international adoption

(Feldstein 2023). Ulnicane aims to answer questions regarding the development of the EU's AI policy and its positioning in comparison to other countries. To achieve this, she contrasts the concepts of 'Normative Power Europe' (Manners 2002) and the 'Market Power Europe' (Damro 2012). The author notes that the EU seeks to promote core norms such as peace and human rights, while leveraging its expensive single market to exert global influence. Additionally, she highlights the EU's contributions, including its endorsement of the OECD's AI Ethics principles, which demonstrate its commitment to international collaboration (Ulnicane 2022).

Justo-Hanani (2022) analyzes the development of the EU's regulatory policy on AI from 2017 to 2021. The research aims to determine the most suitable conceptual approach for understanding the EU's dedication to safeguarding consumer protection and fundamental rights. Justo-Hanani tests multiple hypotheses through three theoretical frameworks: economic competitiveness, institutional structure, and the policy preferences of domestic actors. This approach allows for the identification and evaluation of different stages of the policy-making process (Justo-Hanani 2022).

The EU is also compared to other countries in AI-related research. In a comparison of national AI strategies, Roberts et al. (2023) finds that China prioritizes innovation and 'common prosperity'. The EU's focus is on ethical outcomes and the protection of fundamental rights. According to Roberts et al., this offers the opportunity to learn from and adapt to each other (Roberts et al. 2023).

This overview of various studies has shown that the EU is a popular research topic in the field of AI Ethics and has significant international influence. However, there are still gaps that this research aims to address.

The impact of the EU's AI approach on the US is an area that is currently underrepresented in literature. Birchfield (2022), Greenleaf (2021) and Feldstein (2023) have commented on the EU's *potential* to influence the international AI field

through the AI Act. Hence, their research is especially focused on trying to predict future developments. This is speculative, though, as the AI Act is yet to be adopted by the EU. Moreover, this focus is general and not specifically on the influence on one specific country. In this context, Roberts et al. (2023) compared the EU's and China's AI strategies, while Justo-Hanani (2022) analyzed the EU's policy-making process over time. However, a nuanced comparative analysis of the evolutionary approach to AI Ethics of both the EU and the US over time is missing. Understanding this relationship is essential, given the central role of the US in global AI developments. Moreover, although there is a general consensus that the policies of the EU may have global effects, it is still unclear *how* such influence occurs. For example, Roberts et al. (2023) provided a descriptive comparison of the AI strategies of the EU and China. The mechanisms of diffusion, which refer to *how* AI-related policies and approaches spread as well as influence each other, are not yet fully comprehended in the context of AI. Feldstein (2023) took initial steps in this area, but this study proposes a more detailed examination using process tracing and specific mechanisms based on diffusion theory. This could potentially establish causal diffusion mechanisms between global AI players, specifically in this case the EU and the US.

In summary, this study aims to address three research gaps. Firstly, it will provide a focused comparison between AI Ethics approaches of the EU and the US. Secondly, it will analyze the historical perspective instead of a forward-looking approach. Lastly, it will examine the causal diffusion mechanisms concerning AI Ethics. The research question, 'What is the influence of the EU's AI Ethics approach on the US?' aims at addressing these research gaps. The following chapter will provide a detailed explanation of how the research question will be addressed.

3 Research Design

3.1 Theoretical Framework

This chapter provides the theoretical framework for investigating the influence of the EU's approach to AI Ethics on the US. The theories of Constructivism, Normative Power Europe and Diffusion will be explored in detail, highlighting their relevance and applicability to this thesis.

3.1.1 Constructivism

Constructivism offers a perspective to observe and understand international relations not only as primarily based on material power, but also by emphasizing the significance of norms, beliefs, and ideas in defining state behavior. This is the core of constructivist theory: the assumption that the structure of international politics is more of a social nature than strictly material based. According to Wendt, the structure of the international system is not predetermined but rather continuously produced and reproduced by the actions of its actors (Wendt 1999). As a result, the identity and interests of these actors are defined by this structure, and are deeply interconnected through shared meanings and social values (Wendt 1992, 1999). One of the core elements of constructivism is the role of norms in shaping state behavior. Norms are defined as 'shared expectations about appropriate behavior given a particular identity' (Finnemore and Sikkink 1998). Such norms are not just behavioral regularities but also carry a quality of 'oughtness' and are tied to a particular identity (Finnemore and Sikkink 1998). Constructivist scholars aim to comprehend how political actors establish and disseminate these shared understandings of norms. Wendt argues that norms play a crucial role in shaping states' perceptions of threats and opportunities, and, consequently, their actions on the international stage (Wendt 1999).

The *Life Cycle of Norms*, developed by Finnemore and Sikkink (1998), outlines the evolution of norms in international relations through three stages: norm emergence, norm cascade, and norm internalization. Initially, 'norm entrepreneurs' promote new norms and try to gain acceptance from a critical mass of state leaders. During this phase, international organizations play a critical role as a platform for norm entrepreneurs. Once a tipping point is reached (about 1/3 of the actors accept the norm), the norm cascade phase begins, leading to broad acceptance of the norm due to pressures such as the desire for international legitimacy and conformity. Finally, during the norm internalization phase, these norms become so universally accepted that they are taken for granted (Finnemore and Sikkink 1998).

According to a constructivist perspective, persuasion is the primary mechanism for constructing and reconstructing these social norms. The persuasive power of a norm lies in its ability to outweigh competing norms and interests. Framing plays a crucial role in persuasion by providing a meaningful context that guides the audience towards the desired response or behavior. Frames not only describe the world but also offer ways to act within it (Payne, 2001). Rules, standards, and principles are embedded in norms on different levels. Rules provide clear behavioral expectations based on set conditions, while standards require a post-hoc evaluation of behavior based on broader criteria or underlying policies. They do not suggest a specific action but set a benchmark for evaluating how appropriate it is. Principles are the most general of the three categories, providing broad guidelines for future actions without suggesting a definite normative outcome. They offer guidance, while leaving room for interpretation. Together, these three categories help in understanding the depth or specificity of norms (Finnemore and Hollis 2016).

3.1.2 Normative Power Europe

In the discourse of constructivism and norms, the concept of 'Normative Power Europe' (NPE) emerged. Ian Manners initiated this concept to underscore the EU's

role as a transformative force capable of shaping global politics' conceptions of 'normal,' rather than just an economic or political actor (Manners 2002). The origin of NPE can be traced back to the concept of 'civilian power'. Bull (1982) conceptualized this idea, suggesting that some international entities, such as the EU, exert influence not only through military means but also through economic and diplomatic instruments. The EU's normative power is rooted in its history and general mode of global interaction. After World War II, the EU was formed with the aim of reshaping Europe through collaboration and unity. In the process, a strong commitment to values such as peace, liberty, democracy, the rule of law, and human rights was put. These values not only shape its internal workings but also influence its external interactions, defining its unique identity in the global arena (Manners 2002).

The concept of NPE is centered around various diffusion mechanisms of norms. The uniqueness of the EU may inspire emulation through *contagion* in other parts of the world. Through *informational diffusion*, the EU communicates its principles during diplomatic and multilateral conversations, subtly influencing external agendas. This influence is further enhanced by *procedural diffusion*, which highlights the EU's inherent normative tendencies during external interactions. Moreover, in its external relations, whether it be trade, aid, or strategic partnerships, there is a *transference* of its core values, often accompanied by terms that encourage alignment with its norms. Although *overt diffusion* via sanctions, for example, is less frequent, it remains a tool in the EU's arsenal. It is important to recognize that the *cultural filter* - how external actors interpret and adopt EU norms - is deeply affected by the respective actor's individual historical and cultural context which has an influence of the effectiveness of the EU's normative power. Countries that have historical ties with the EU and share similar values or political systems are more likely to be receptive to the EU's normative messages (Manners 2002).

3.1.3 Diffusion

The concept of diffusion is central in understanding how and why policies, ideas, norms, and innovations flow across countries and regions. As Rogers et al. define it, diffusion is generally "the process by which an innovation is communicated through certain channels over time among the members of a social system" (Rogers et al. 2009). Political diffusion is characterized by decisions in one country being systematically influenced by prior policy choices made in other countries. This interdependence allows for a focus on the specific mechanisms that lead to the spread of policies, rather than the outcome (Gilardi 2013). Generally, literature identifies four broad categories of diffusion mechanisms: coercion, competition, learning, and emulation (Gilardi 2013).

Competition is defined as "the process whereby policy makers anticipate or react to the behavior of other countries in order to attract or retain economic resources" (Gilardi 2013). The focus is clearly on economic benefits. However, this can lead to a downward spiral, where countries lower their standards or regulations to attract business (Gilardi 2013). It is important to note that this is not always the case. An open economy does not necessarily result in fewer domestic regulations (Elkins et al. 2006; Vogel 2009). On the contrary, countries may increase their domestic standards in response to international market pressures (Vogel 2009). Competitive diffusion often leads to the introduction of standards, principles, and laws in order to seemingly adhere to international standards. However, this does not guarantee implementation. The term 'ethics shirking' or 'ethics bluewashing' (Cao and Prakash 2012) was previously used to describe this phenomenon. Gilardi defines learning as "the process whereby policy makers use the experience of other countries to estimate the likely consequences of policy change" (Gilardi 2013). Policy makers hold certain beliefs about policy outcomes, but when presented with evidence from other countries (both positive and negative in nature), these beliefs may change, leading to corresponding changes in action. In this context, it has been

observed that countries are more likely to adopt policies that are perceived as successful in other countries (Meseguer and Gilardi 2009; Elkins et al. 2006). A potential problem in the context of learning is cognitive bias and subjectivity (Weyland 2009). This bias can be strongly influenced by ideological beliefs, which can lead to selective learning (Shipan and Volden 2008; Gilardi 2010). Learning from other countries is a common diffusion mechanism. However, it may not always serve the public interest of the adopting country objectively due to bias problems.

Emulation refers to "the process whereby policies diffuse because of their normative and socially constructed properties instead of their objective characteristics" (Gilardi 2013). In the case of emulation, political actors decide based on 'appropriateness' as opposed to 'consequences' (Checkel 2005; Gilardi 2013). Therefore, emulation is less about outcome than other diffusion mechanisms. Emulation is primarily concerned with norms and they typically follow the *Life Cycle of Norms*: the emergence, its widespread acceptance, and its eventual universal internalization of a norm (Finnemore and Sikkink 1998). It is a challenge to fully understand the norms in question; therefore, a qualitative analysis is often needed (Weyland 2009; Brooks 2005).

Coercion in the international context refers to powerful entities pressuring states to adopt certain policies through a mechanism called 'conditionality' (Gilardi 2013). For instance, international financial institutions often tie financial aid to neoliberal reforms (Biersteker 1990). Similarly, the EU requires broad reforms and the adoption of EU laws for new members (Schimmelfennig and Sedelmeier 2004). However, the effectiveness of such strategies is debated. Some studies suggest limited success in achieving desired reforms (Brooks 2005; Weyland 2009). Bradford's concept of the 'Brussels Effect' is a mixture of various forms of diffusion mechanisms. Bradford describes how the EU promotes de facto influence by requiring global companies to adhere to EU regulations in the European market, as well as de jure influence by non-EU jurisdictions to adopt regulations similar to

those of the EU. Bradford (2012) suggests that when non-EU jurisdictions lack legislation on a particular topic, they are more likely to emulate the EU's legislation. The 'Brussel's Effect' therefore represents a combination of competitive, learning, and emulative diffusion mechanisms (Bradford 2012).

3.1.4 Combined Theoretical Framework

This chapter outlines the benefits of individual theories for this research and explains why a combined theoretical framework is deemed useful for the research purpose.

Constructivism emphasizes the importance of shared norms, ideas, and beliefs in the international system, making it a suitable lens for this research. As outlined in the literature review, AI governance and its respective challenges require states to enhance their international cooperation. This means, that individuals will need shared norms, ideas, and beliefs. Constructivism and in particular the Life Cycle of Norms help us to understand how these shared AI-related beliefs and norms emerge and evolve within the international system.

The Normative Power Europe concept assumes that the EU, based on a strong commitment to values and human rights, has a unique international position with regards to influencing what is perceived to be 'normal'. This shall be an underlying assumption of this research, too. Hence, this research assumes that the EU is uniquely capable of influencing what is perceived to be 'normal' in the realm of AI Ethics.

The concept of diffusion provides insights into *how and why* ideas, policies and norms move within the international system. To better understand the impact of the EU's approach to AI Ethics on the US, it may be useful to operationalize the diffusion mechanisms of coercion, learning, emulation, and competition. This could provide insights into the causal relationships between the two.

As this research is multilayered, a combined theoretical framework can be beneficial. The interplay of the theories is particularly relevant when addressing AI Ethics related questions. This area is not solely concerned with technological advancement but is deeply embedded in societal values, beliefs, and standards. Understanding the creation of norms (Constructivism), the unique normative influence of the EU (NPE), and the processes that allow these norms to spread (Diffusion) provides a natural flow and equips this research with a comprehensive combined theoretical framework.

3.2 Hypotheses

The theoretical framework discussed provides potential underlying assumptions for the research to explore. In order to do so, in the empirical analysis, the research will test a variety of hypotheses embedded within the theoretical framework.

H1: The EU perceives itself as a norm entrepreneur with regards to AI, actively promoting its ethical approach in the international system.

Constructivism emphasizes the role of norms, ideas, and beliefs within the international system. The Life Cycle of Norms suggests that norm entrepreneurs attempt to promote their beliefs globally in the beginning of a norm's life cycle. The Normative Power Europe concept assumes that the EU inherently acts as a norm entrepreneur. These assumptions are applied to AI Ethics in H1.

H2: The US adopts a similar approach to AI Ethics due to coercive pressure from the EU.

The diffusion mechanism coercion is rooted in a mechanism referred to as conditionality. When looking at the AI Ethics relationship between the EU and the US, it is possible that diffusion is at play. Therefore, it will be investigated through H2.

H3: The US adopts a similar approach to AI Ethics based on learning from the EU's experiences and frameworks.

This hypothesis refers to the diffusion mechanism of learning. In the case of AI Ethics, the US could perceive the EU's approach as successful and therefore worth imitating. This potential mechanism will be tested via H3.

H4: The US emulates a similar approach to AI Ethics as the EU's due to a perception of appropriateness.

In the case of emulation, political actors base their decisions on appropriateness rather than a specific outcome. In the case of AI Ethics, the US could perceive the EU's approach to AI Ethics as appropriate and therefore decide to shift towards it. This potential mechanism will be tested through H4.

H5: The US adopts a similar approach to AI Ethics to the EU's in response to competitive pressure.

The competition diffusion mechanism assumes that countries adopt policies from other countries in order to maximize economic benefits. In this research, the US might perceive as economically beneficial to adopt a similar approach to AI Ethics as the EU's. Therefore, this potential mechanism will be tested via H5.

3.3 Methodology

3.3.1 General approach

When studying diffusion, researchers must decide whether to use quantitative or qualitative research methods. Both approaches offer advantages and challenges, especially in capturing the mechanisms of diffusion processes. Quantitative methodologies are useful for identifying the presence and strength of diffusion processes across multiple countries. They can systematically capture the existence,

relevance, and general direction of diffusion. Statistical techniques can be used to analyze the facilitating or hindering factors of diffusion, particularly those of temporal nature (Jahn 2022). Moreover, quantitative approaches, when based on clear theoretical frameworks and careful empirical testing, can provide valuable insights into social phenomena (King et al. 2021). However, while quantitative methods are effective at identifying the 'what' and 'how much' of diffusion, they are weaker at identifying the underlying causal mechanisms. This is where qualitative approaches excel, as they can zoom into specific scenarios and explore nuanced causal relationships (Starke 2013). They emphasize the micro-level explanations of individual actors, making it better suited to understand underlying causal mechanisms and differentiate them from correlation (Gerring 2005; Hedström and Ylikoski 2010). Therefore, a qualitative approach is selected to identify the effect of the EU's AI ethics approach on the US and underlying diffusion mechanisms, which requires a causal relationship. In this context, process tracing is a particularly suitable method (Jahn 2022; Starke 2013).

3.3.2 Process Tracing

Process tracing is a powerful method to trace and understand the causal sequences and diffusion mechanisms. Collier defines it as “the systematical examination of diagnostic evidence selected and analyzed in light of research questions and hypotheses posed by the investigator” (Collier 2011). In his framework for process tracing, Collier differentiates between three distinct characteristics of process tracing. First, process tracing focuses on evidence to infer causality. This evidence is often referred to as causal-process observations (CPOs). Second, while process tracing seeks to unravel causal mechanisms, a careful description remains highly important as it serves as the foundation for making causal observations. Third, sequences of independent, dependent, and intervening variables are a core component of process tracing (Collier 2011). At its core, process tracing aims to determine why and how (evidence) certain events (description) evolved over time

(sequences). In many studies, the causal link between an independent variable and its respective outcome remains unclear. This is often referred to as the 'black box' of causality (Trampusch and Palier 2016). In process tracing, 'causal mechanisms' are conceptualized as a continuous sequence of cause and effect, initiated by entities, linking a possible cause to its anticipated result and thereby 'opening' the black box of causality (Beach and Pedersen 2019). Beach and Pedersen (2019) distinguish between four distinct types of process tracing, each of which serve a unique function in research. All types require knowledge about an outcome.

- Theory-testing Process-Tracing: In this case, hypothetical causal mechanisms are conceptualized based on existing theory and empirical research. Research collects and assesses empirical evidence in order to evaluate if the expected mechanisms are present and function according to the tested theory.
- Theory-Building Process-Tracing: In this case a cause (or a set of causes) is tried to be connected to a given outcome without prior knowledge of potential linking mechanisms. Therefore, the goal is to identify these mechanisms.
- Theoretical-Revision Process-Tracing: This type is concerned with mechanisms in deviant cases. It tries to answer why mechanisms that should have been in place but did not work out and thereby identify potentially hidden conditions for the respective mechanisms to work out.
- Explaining-Outcome Process-Tracing: This seeks to explain the causal mechanism in a specific case where the outcome is already known, whether empirical or theoretical. The goal is to find answers to puzzling outcomes or theoretically unexpected outcomes.

As previously mentioned, process tracing aims to identify causality, and description plays an important role in achieving this goal. To comprehend the sequence of events, each step must be precisely described. This descriptive sequencing can then be used to construct a detailed timeline that chronologically captures the events relevant to the research. The timeline and sequences can be used to trace and analyze

the causal mechanisms that lead to the examined outcome. This can help identify critical junctures or turning points (Collier 2011).

Qualitative evidence tests were first introduced to prove causality in general (van Evera 1997) and later adopted for process tracing (Collier 2011; Mahoney 2012). They are an essential tool for identifying causal mechanisms between sequences of events and working towards explaining outcomes. Their different layers help to systematically evaluate the validity and strength of the evidence. This helps to open the black box of causality and thereby accept or deny the respective hypothesized causal mechanisms. There are four levels of evidence tests (van Evera 1997; Collier 2011; Mahoney 2012):

- Straw-in-the-Wind Tests: With the help of this test, the plausibility of a hypothesis can either be increased or decreased, but they are not decisive on their own due to weak evidence. Therefore, they do not provide necessary or sufficient criteria for accepting or rejecting a hypothesis. In the example 'X murdered Y', evidence for passing this test could be that X has a motive to murder Y. This test alone is not sufficient to accept or reject a hypothesis. It is the weakest of the four tests. However, if a hypothesis passes multiple straw-in-the-wind tests, it accumulates significant evidence.
- Hoop Tests: As the name already suggests, a hypothesis must 'jump through the hoop', hence meet a specific criterion to pass the test. When evidence fails to meet this criterion, the corresponding hypothesis can be ruled out. In the given example, the absence of an alibi for X would serve as evidence for passing the test. Although passing the test does not entirely confirm the hypothesis, the necessary criteria for the hypothesis to be true are met. Compared to straw-in-the-wind tests, passing hoop tests has stronger implications, because it significantly weakens the plausibility of rival hypotheses.

- Smoking-Gun Tests: As the name suggests, this test searches for strong and direct evidence, similar to finding a suspect with a smoking gun at the crime scene. In the given example, this would mean that X is found with a smoking gun at the crime scene. If this is the case, it provides a sufficient but not necessary criterion for accepting the hypothesis. Therefore, the given hypothesis is strongly supported. However, failure to find such evidence does not necessarily mean that the hypothesis is rejected.
- Doubly Decisive Tests: To pass this test, the evidence must align perfectly with the hypothesis, making it both necessary and sufficient for the hypothesis to be true. For the given example, this could mean a video recording clearly identifying X as the murderer. In this case the hypothesis would fully confirm, and eliminate all other possibilities. However, such evidence is rare in social science.

This analysis will only perform hoop tests, smoking-gun tests, and doubly decisive tests. Straw-in-the-wind tests are excluded as they do not provide sufficient evidence for causal inference. Although hoop tests also do not meet the criteria for causal inference, they are useful in setting the criteria for seeking evidence to support the respective hypothesis. Moreover, in the context of this research, hoop tests serve a very specific function that will be explained in the following chapter.

3.4 Operationalizing the Hypotheses

To answer the research question, 'What is the effect of the EU's AI Ethics approach on the US?', this empirical analysis of this research is divided into three stages. The first stage provides a systematic overview of the EU's approach to AI Ethics over time to establish an empirical basis. In this process, the research aims to validate H1 and provide evidence that the EU perceives itself as a global norm entrepreneur in the field of AI Ethics. In the second stage, a systematic analysis of the US's approach to AI Ethics will be performed (mirroring the previous analysis of the EU for

comparability). Based on both analyses, in the third stage, this research will dive into H2-H5 and aim to a) identify *if* the US's approach to AI Ethics has shifted towards the EU's approach. If such a shift has occurred, this research b) will identify potential diffusion mechanisms that explain *why*. This research is comparative in nature. It is situated within an X-centered research design. The goal is to determine the effect the explanatory variable (the EU's approach to AI Ethics) has on the dependent variable (the US's approach to AI Ethics). The literature review highlights the EU's global influence in the field of AI Ethics and its unique normative role within. Moreover, as the Normative Power Europe concept demonstrates, the EU has a unique normative role within the international system. Therefore, this research design chose the EU as a case and its approach to AI Ethics as the explanatory variable. The US was also chosen due to a) its close ideological proximity (and therefore a higher expectation of diffusion (Gilardi 2013; Manners 2002)) and b) its important global role with regards to AI. The US currently holds the top spot in the most recent Global AI Index, which ranks countries based on investment, innovation, and implementation criteria (Tortoise 2023).

Stage 1: Empirical analysis of the EU's approach to AI Ethics

H1: The EU perceives itself as a norm entrepreneur with regards to AI, actively promoting its ethical approach in the international system.

The first stage has two objectives. A) Validating the basic assumption of the H1, which is based on the concept of Normative Power Europe, that the EU seeks to influence other countries regarding AI Ethics and perceives itself as a respective norm entrepreneur. B) As outlined in the process tracing chapter, it is essential to capture and describe relevant sequences for the research. The second objective is to provide a detailed overview of the EU's approach to AI Ethics over time. This is crucial because it creates the empirical foundation for comparing the US approach to AI Ethics. To create this foundation, this empirical foundation will conduct a

systematic analysis of official EU AI-related documents. The selection criteria for these documents are:

- Date of publication: January 2018 – August 2023
- Source: https://digital-strategy.ec.europa.eu – this is the digital repository of official digital-related documents of the EU
- Type of document: Official policy document that specifically has AI as a topic, not directed at a specific sector or field but at AI in general; only EU-level documents, no national documents

These documents are then analyzed through 'skimming (superficial examination), reading (thorough examination), and interpretation' (Bowen 2009). Relevant text passages will be identified in the process based on pre-defined codes. The following codes will be used in the analysis:

Table 1: Codes and categories for the analysis of EU documents

Code name	*Category*
E1	Any ethical principle/value that is mentioned
E2	Any AI-related risk/challenge that is mentioned
E3	Recommended AI Ethics related action
E4	AI Ethics related definition
E5	Self-perception of the EU as a potential AI Ethics norm-entrepreneur
E6	AI Ethics related coercive pressure of the EU

Source: author's own.

E1-E4 seek to capture the different facets of the EU's approach to AI Ethics. E5 specifically aims to validate H1. E6 serves to gather evidence for H3. The results are presented in chronological order, which is in line with the logic of process tracing which evaluates evidence over time.

Stage 2: Empirical analysis of the US's approach to AI Ethics

The second stage is a mirrored analysis of the US's approach to AI Ethics. The objective here is to provide an overview of the US's approach to AI Ethics over time. This is essential as it provides the empirical basis for sequences against which the EU's approach to AI Ethics can be compared. Consequentially, a qualitative document analysis of relevant US documents will be performed, following the same structure as the analysis of the EU documents. The following document selection criteria will be applied:

- Date of publication: January 2018 – August 2023
- Source: https://ai.gov – this is the digital repository of official AI-related documents of the US
- Type of document: Official policy document that specifically has AI as a topic, not directed at a specific sector or field but at AI in general; only federal-level documents, no state documents

These documents will then also be analyzed through 'skimming (superficial examination), reading (thorough examination), and interpretation' (Bowen 2009). Relevant text passages will be identified in the process based on pre-defined codes. As shown in Table 2, the following codes will be used in the analysis.

U1-U4 reflect the codes used in the analysis of EU documents and seek to capture different facets of the US's approach to AI ethics. U5 aims to identify potential references to the EU that could serve as evidence for H2-H5. U6 will gather evidence for H4, U7 for H5 and U8 for H3. Mirroring the analysis of the EU, the findings will be presented chronologically for comparability.

Table 2: Codes and categories for the analysis of US documents

Code name	*Category*
U1	Any ethical principle/value that is mentioned
U2	Any AI-related risk/challenge that is mentioned
U3	Recommended AI Ethics related action
U4	AI Ethics related definition
U5	Mention of the EU
U6	Mention of international cooperation
U7	Any mention that highlights competitive pressure
U8	Any mention that highlights coercive pressure

Source: author's own.

Stage 3: Analysis of the influence of the EU's approach on the US and potential diffusion mechanisms

The objective of the third step is twofold. A) It seeks to identify *if* the US's approach to AI Ethics has shifted toward the EU approach, and b) if that is the case, it seeks to identify potential diffusion mechanisms that might explain why this shift has occurred. This raises the problem that, at the beginning of this stage, the outcome is not yet specified. However, each of the four types of process tracing aims to analyze cause-outcome effects, which requires prior knowledge of the outcome. This establishment of the outcome is in line with objective a). In order to do so, the US approach to AI Ethics will be closely compared with that of the EU over time. In this context, a detailed timeline will be constructed that provides an overview of the chronology of both the EU's and the US's approaches to AI Ethics. The establishment of the potential outcome (shift of the US's approach to AI ethics towards the EU's approach to AI Ethics) will be tested for each hypothesis using a hoop test. If the hoop test is passed and a shift in the US approach to AI Ethics towards the EU approach is observed, the analysis will move on to the second

objective of stage three – identifying potential diffusion mechanisms that could explain this shift. At this point in the research the explanatory variable is no longer the EU's approach to AI Ethics anymore. The question is: Does X (explanatory variable: coercion, learning, emulation, competition) cause X (the potential shift from the US's approach to AI Ethics to the EU's approach)? By operationalizing the four broad mechanisms of diffusion via H2-H5, process tracing is theory-testing, as it seeks to test the validity of the diffusion mechanisms in the context of a given outcome. The objective is to establish causal diffusion mechanisms via smoking-gun and/or doubly decisive tests.

The testable evidence will be based on the document analysis of both the EU's and the US's approaches to AI Ethics. Process tracing is an iterative process that requires some flexibility. If potential new evidence emerges during the document analysis, this evidence may need to be addressed additionally. An example would be joint declarations or ethical principles of an independent international organization that have been signed by both parties. This new evidence would again be analyzed via qualitative document analysis.

The following evidence tests will be applied for each hypothesis:

H2: The US adopts a similar approach to AI Ethics due to coercive pressure from the EU.

- Hoop Test: The US changes its approach to AI Ethics, moving closer to the EU's approach.
- Smoking-Gun Test: The US changes its AI Ethics approach and thereby moves towards the EU's approach, shortly after coercive pressure from the EU.
- Doubly Decisive Test: Direct statements in official documents confirming alignment with the EU's AI Ethics approach in response to EU coercion.

H3: The US adopts a similar approach to AI Ethics based on learning from the EU's experiences and frameworks.

- Hoop Test: The US changes its approach to AI ethics, moving closer to the EU's approach.
- Smoking-Gun Test: The US changes its approach to AI ethics, moving toward the EU's approach, shortly after citing the EU's successful approach to AI ethics.
- Doubly Decisive Test: Direct statements in official US documents confirming the adoption of a similar approach to AI Ethics based on learning from the EU's experiences and frameworks.

H4: The US emulates a similar approach to AI Ethics as the EU's due to a perception of appropriateness.

- Hoop Test: The US changes its approach to AI ethics, moving closer to the EU's approach.
- Smoking-Gun Test: The US makes changes to its AI Ethics, thereby moving towards the EU's approach, and justifies the change on the basis of appropriateness.
- Doubly Decisive Test: Direct statements in official US documents confirming alignment with the EU's AI Ethics approach for appropriateness.

H5: The US adopts a similar approach to AI Ethics to the EU's in response to competitive pressure.

- Hoop Test: The US changes its approach to AI ethics, moving closer to the EU's approach.
- Smoking-Gun Test: The US makes changes to its AI Ethics and thereby moves towards the EU's approach, shortly after acknowledging the competitive benefits of promoting ethical AI principles.
- Doubly Decisive Test: Direct statements in official US documents confirming the adoption of an AI Ethics approach similar to the EU's in response to competitive pressure.

3.5 Limitations of the Research Design

A key limitation is the full reliance on documents as a source throughout the analysis. While this thesis seeks to provide robustness through a structured and focused comparative approach, it must be acknowledged that important evidence could be missed by relying solely on documents. For example, important documents may be missed during the document scoping phase. It is also very likely that additional evidence exists outside of the documents. This is especially the case in diffusion analysis, as it is strongly about the underlying motivations of political actors, which are not always reflected in officially released documents. Moreover, qualitative approaches always leave room for interpretation and, thus, for interpretative bias. Different individual backgrounds, prior knowledge, or research goals may lead to different interpretations. As noted above, the complex concepts of AI Ethics can be challenging in this context. Drawing direct links between observed data and abstract terms may not always be straightforward. While the evidence tests do add robustness, they are inherently qualitative in design and interpretation due to the subjectivity of the author, and therefore also prone to bias.

Another limitation is the focus on the EU and the US. Because of this focus, other potential influences on the US's approach to AI Ethics may be overlooked. In addition, the research findings are not necessarily generalizable. Moreover, while a chronological timeline helps to understand the sequence of events, it may unintentionally overemphasize temporal connections. These sequences need to be evaluated carefully, so that causation is not automatically inferred from chronology alone. This is where evidence testing and process tracing can help identify causal mechanisms.

And finally, AI systems are evolving rapidly, and so is AI Ethics. The author acknowledges that this thesis and its findings capture only a specific temporal snapshot of a highly dynamic field. As a result, the conclusions may have limited applicability as the field evolves.

4 Empirical Analysis

As outlined previously, the empirical analysis of this research is divided into three stages. The first stage involves a systematic analysis of the EU's approach to AI Ethics over time. The objective is to validate H1 and provide evidence that the EU perceives itself as a global norm entrepreneur in the field of AI Ethics. It will also provide the empirical basis for a comparative analysis in the third stage. In the second stage, a systematic analysis of the US approach to AI Ethics will be performed (thereby mirroring the previous analysis of the EU for comparability). Stage three dives into H2-H5 and seeks to a) identify *if* the US's approach to AI Ethics has shifted toward the EU approach and, if so, b) identify potential diffusion mechanisms that explain *why* such a shift has occurred.

4.1 The EU's approach to AI Ethics

This chapter conducts the first stage of the empirical analysis. This includes an analysis of a) how the EU's approach to AI Ethics has evolved over time, and b) what the self-perception of the EU is in terms of AI Ethics and respective AI Governance. To this end, the analysis will closely follow the operationalization as outlined in the previous chapter. As a first step, the following documents were identified based on the selection criteria as seen in Table 3.

Next, the document analysis method of 'skimming, reading, and interpretation' Bowen (2009) will be applied. During skimming and reading, relevant text passages are categorized according to the coding scheme in Figure 1. The codes help to highlight the EU's ethical principles, AI concerns, proposed ethical AI actions, how the EU defines AI Ethics terms, and its view of its global role in AI Ethics. The findings of these text passages will then be filtered and interpreted.

The results of this interpretation will be presented in chronological order, following the logic of process tracing, which evaluates evidence over time.

Table 3: Documents selected for the analysis of the EU's approach to AI Ethics

No.	*Title*	*Date*
1	Artificial Intelligence for Europe (European Commission 2018a)	25.04.2018
2	Coordinated Plan on Artificial Intelligence (European Commission 2018b)	07.12.2018
3	Ethics Guidelines for Trustworthy AI (AI HLEG 2019a)	19.04.2019
4	Policy and Investment Recommendations for Trustworthy AI (AI HLEG 2019b)	26.06.2019
5	White Paper: On Artificial Intelligence – A European approach to excellence and trust (European Commission 2020)	19.02.2020
6	Coordinated Plan on Artificial Intelligence 2021 Review (European Commission 2023a)	21.04.2021
7	Laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts (European Commission 2021)	21.04.2021

Source: author's own.

Due to the large amount of information gathered during the skimming and reading phase, it is located in the appendix of the thesis[1]. Each relevant text passage

[1] The passage numbers mentioned can be found in the appendix, which can be accessed via the following link: https://ib.uni-koeln.de/sites/jaeger/publikationen/aipa/8_Appendix_AIPA_12024.pdf.

is listed, numbered, and categorized. For traceability and transparency, this analysis will refer to passage numbers in square brackets throughout the analysis.

Document analysis of the EU's approach to AI Ethics

The European Commission's (EC) Communication 'Artificial Intelligence for Europe' (April 2018) serves as the initial specifically AI-related strategic document issued by the EU. It is essentially two-fold in nature: first, it highlights the need for the EU to be competitive in the realm of AI. It urges increased investment in AI research and innovation, calls for a coordinated approach among members, and emphasizes the significance of preparing Europe's workforce for an AI-driven future. It also highlights on several occasions that AI must be developed and deployed based on European values [2, 3, 8, 17, 22, 23] and fundamental rights [2, 3, 8, 10, 14, 22] due to the multifaceted risks that AI can pose [11, 12]. Trust emerges for the first time as a key principle [7, 10, 11]. The Explainability [11] of AI is identified as a precondition for trust and is therefore also highlighted. Moreover, the document introduces the concept of a human-centric [4] approach to AI. Several other principles necessary for the development and deployment of AI are mentioned throughout the document: safety [12, 14, 15], security [14, 15], inclusion [4, 5, 14], transparency [2, 11, 14], non-discrimination [5, 14], privacy, dignity, and fairness [14]. However, none of these principles/values are specifically defined. This will be the task of draft AI Ethics guidelines to be developed [3, 13, 14]. Regarding the self-perception of the EU as a global champion of AI Ethics, the document provides first evidence: the EU clearly positions itself as a potential international norm entrepreneur for AI Ethics based on its values [17-23].

This self-perception is further emphasized in the 'Coordinated Plan on Artificial Intelligence' (December 2018), the next AI-related strategic document published by the EC. Here, it clearly states that the EU intends to promote an ethical

approach to AI globally [35, 36, 38]. The document serves as a call for a coordinated approach to AI among EU member states in order to be internationally competitive. Investment, research and innovation, skills and education, as well as data and infrastructure are highlighted as essential for competitivness. Moreover, the term 'Trustworthy AI' [25] is introduced, where trust is defined as being predictable, responsible, verifiable, respecting fundamental rights, and following ethical rules [30]. Together with human-centric AI, it is affirmed as central to the development of AI [24, 35]. In addition, this document suggests that an AI specific regulatory framework may be needed to 'promote innovation while ensuring high levels of protection and safety' [33].

The 'Ethics Guidelines for Trustworthy AI' (April 2019), produced by the EC appointed High-Level Expert Group on Artificial Intelligence (AI HLEG), serves as a foundational document for the EU's understanding of AI Ethics and introduces multiple definitions of relevant terms. 'Trustworthy AI' is identified as the 'foundational ambition' [41] of AI systems and is broken down into three necessary but not sufficient components [45]:

- lawful (compliance with applicable law and regulation)
- ethical (adherence to ethical principles and values)
- robust (technically and socially)

The guidelines focus on the ethical and robust components. Ethical AI is 'used to indicate the development, deployment, and use of AI that ensures compliance with ethical norms, including fundamental rights of specific moral entitlements, ethical principles, and related core values' [98]. Robust AI is described as follows:

> Robustness of an AI system encompasses both its technical robustness (appropriate in a given context, such as the application domain or life cycle phase) as well as its robustness from a social perspective (ensuring that the AI system duly takes into account the context and environment in which the system operates). This is crucial to ensure that, even with good intentions, no unintentional harm can occur. [100]

Further, it lists four ethical principles necessary for a trustworthy development, deployment, and use of trustworthy AI [66]:

- Respect for human autonomy
- prevention of harm
- fairness
- explicability

Human-centric AI is at the core of respecting human autonomy and is defined as:

> The human-centric approach to AI strives to ensure that human values are central to the way in which AI systems are developed, deployed, used and monitored, by ensuring respect for fundamental rights, including those set out in the Treaties of the European Union and Charter of Fundamental Rights of the European Union, all of which are united by reference to a common foundation rooted in respect for human dignity, in which the human being enjoy a unique and inalienable moral status. This also entails consideration of the natural environment and of other living beings that are part of the human ecosystem, as well as a sustainable approach enabling the flourishing of future generations to come. [99]

Prevention of harm includes the protection of human dignity and integrity, safe, secure, and robust AI systems, and a special focus on potentially vulnerable persons [68]. Fairness consists of equal and fair distribution of benefits and costs, freedom from unfair bias, discrimination, and stigmatization, equal access, and freedom of choice. Furthermore, it includes the ability to seek redress against AI based decisions [69]. Explicability is strongly based on the transparency of the respective AI system in terms of its capabilities, purpose, and decision-making [70].

The guidelines translate these ethical principles into seven actionable requirements, each which is being described in detail [72-93]:

- Human agency and oversight, including fundamental rights, human agency, and human oversight.
- Technical robustness and safety, including resilience to attack and security, fall back plan and general safety, accuracy, reliability, and reproducibility.
- Privacy and data governance, including respect for privacy, quality and integrity of data, and access to data.

- Transparency, including traceability, explainability and communication.
- Diversity, non-discrimination, and fairness, including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation.
- Societal and environmental well-being, including sustainability and environmental friendliness, social impact, society, and democracy.
- Accountability, including auditability, minimization and reporting of negative impacts, trade-offs, and redress.

The second deliverable of the AI HLEG was the 'Policy and Investment Recommendations for trustworthy AI' (July 2019), which mainly addresses the first component (lawful) of trustworthy AI and translates the ethical guidelines into actionable policy and investment recommendations. It also provides specific regulatory recommendations based on the guidelines. In this context, a risk-based approach to the regulation of trustworthy AI is introduced [137, 143, 169], according to which risks, and the corresponding regulatory response should be categorized into different classes [143]. Risk is here broadly defined as 'adverse impacts of all kinds, both individual and societal' [143]. Both the ethics guidelines [101-104] and the policy and investment recommendations [170-172] position the EU as a potential international entrepreneur of human-centric and trustworthy AI.

Following these two AI HLEG deliverables, the EC published the 'White Paper on Artificial Intelligence – A European Approach to Excellence and Trust' in February 2020. Building on the recommendations of the AI HLEG, the EC aims to foster an ecosystem of excellence and an ecosystem of trust [179]. It welcomes the seven key requirements for trustworthy AI put forward by the AI HLEG [180] and identifies regulatory gaps regarding transparency, traceability, and human oversight [182]. As a result, it proposes a future regulatory framework to ensure the ecosystem of trust [179]. According to the White Paper, a risk-based approach (as recommended by the AI HLEG) requires a clear distinction between different risk categories of AI systems. In this context, the concept of high-risk AI applications is introduced [205]. Two cumulative criteria lead to an AI application being considered

high-risk – that significant risks can be expected to occur in a particular sector, and that significant risks are likely to arise from its use [206]. Several potential key requirements for high-risk AI applications are presented: training data; data and record-keeping; information to be provided; robustness and accuracy; human oversight; and specific requirements for certain particular AI applications, such as those used for purposes of remote biometric identification [207].

In chapter 1H, the White Paper highlights the role the EU is already playing in influencing international discussions on the ethical use of AI. In particular, it highlights the EU's close involvement in the development of the OECD's Ethical Principles for AI [223]. In April 2021, the EC published two major AI-related policy documents: The 'Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence' (also known as the 'AI Act') and the 'Coordinated Plan on Artificial Intelligence 2021 Review'. The AI Act marks a significant evolution in the EU's approach to AI, as it is the first specifically AI-related legislative initiative proposed by the EC. Deeply rooted in the strategic progressions in earlier documents, it aims to promote 'an ecosystem of trust by proposing a legal framework for trustworthy AI' [255]. Rules for AI should be human-centric in nature, thereby respecting fundamental human rights [255]. In line with the risk-based approach introduced by the AI HLEG and adopted in the White Paper, the AI Act introduces a more nuanced risk-based approach, classifying AI applications into different risk levels: minimal, limited, high and unacceptable risk [277]. Risks are unacceptable if they contradict EU values by violating fundamental rights and are therefore prohibited. Manipulative or exploitative AI systems, AI-based social scoring, and 'real time' remote biometric identification systems for law enforcement fall under this category [277]. The focus of the AI Act is on high-risk AI applications. These will only be allowed to enter the EU market after fulfilling strict compliance

requirements and an ex-ante conformity assessment [278]. The Appendix[2] of the AI Act lists currently identified high-risk AI systems [341-349]:

- biometric identification
- management and operation of critical infrastructure
- education and vocational training
- employment, workers management and access to self-employment
- access to and enjoyment of essential private services and public services and benefits
- law enforcement, migration
- asylum and border control management
- administration of justice and democratic processes

Strictly necessary requirements for mitigating the risk of high-risk AI applications are high quality data, documentation and traceability, transparency, human oversight, accuracy and robustness [255].

In essence, the AI Act is a testament to the EU's commitment to pioneering a balanced approach to AI—one that fosters innovation while protecting individual rights and societal values. However, it is worth noting that the Act is still subject to debate and discussion, with various stakeholders providing feedback. The final form of the regulation may evolve as it undergoes the legislative process in the EU. This legislative effort underscores the EU's ambition not only to regulate AI within its borders, but also to set a global benchmark for AI Ethics and governance [337, 358].

The 'Coordinated Plan on Artificial Intelligence 2021 Review' rather presents strategic considerations for the competitiveness of the EU in the AI sector. It builds

[2] The passage numbers mentioned can be found in the appendix, which can be accessed via the following link: https://ib.uni-koeln.de/sites/jaeger/publikationen/aipa/8_Appendix_AIPA_12024.pdf.

on the first Coordinated Plan on AI of 2018, reflecting the evolving landscape of AI and the lessons learned since 2018. The focus is on how to enable innovation, attract talent, and identifying high impact sectors where the EU can compete. At the same time, it commits to implementing the AI Act. In terms of AI Ethics, nothing new can be observed. However, the EU strongly emphasizes its self-perception as a global leader in human-centric and trustworthy AI [240-247]. In particular, it notes that it co-founded the Global Partnership on AI (GPAI) [242] and formulates the goal of working towards an AI agreement with the US. For this purpose, an EU-US Trade and Technology Council is proposed [243].

Conclusion

The EU's approach to AI Ethics, as evidenced by the documents analyzed, has been a progressive journey from broad policy recommendations to detailed, actionable frameworks. This evolution can be divided into three phases.

Initiation Phase: The initial documents, the 'Communication Artificial Intelligence for Europe' and the 'Coordinated Plan on Artificial Intelligence (2018)', laid the groundwork for the EU's strategic vision on AI. They emphasized European values and fundamental rights, and introduced key terminologies such as trustworthy and human-centric AI, as well as numerous other ethical principles necessary for the development, deployment, and use of AI.

Operational Phase: The 'Ethics Guidelines for Trustworthy AI' and the 'Policy and Investment Recommendations for Trustworthy AI', moved from broad definitions and strategic visions to more granular, actionable frameworks and terminologies. This phase introduced the risk-based approach to AI and assessed the need for an ethics based regulatory framework.

Regulatory Phase: Recent documents, such as the 'White Paper on Artificial Intelligence', the 'Coordinated Plan on Artificial Intelligence 2021 Review', and the

'AI Act', represent a maturation of the EU's approach. They specify potential regulatory measures and introduce a categorical risk-based differentiation of AI applications, introducing the term high-risk AI system.

To fully confirm hypothesis 1 (EU as a self-perceived global leader in AI Ethics) using the Doubly Decisive test, as outlined in the Operationalization chapter, two types of evidence are needed: clear evidence supporting the hypothesis in the documents analyzed and no evidence contradicting the hypothesis in the same documents. The EU's self-perception as a global leader in AI Ethics is consistently emphasized and therefore evident throughout the documents. The 'Communication Artificial Intelligence for Europe' [17-23], the 'Coordinated Plan on Artificial Intelligence (2018)' [35-38], the 'Ethics Guidelines for Trustworthy AI' [101-104], the 'Policy and Investment Recommendations for Trustworthy AI' [170-172], the 'White Paper on Artificial Intelligence' [221-224], the AI Act [237, 258], and the 'Coordinated Plan on Artificial Intelligence 2021 Review' [240-247] all emphasize the EU's ambition to set global standards and influence the worldwide AI debate based on its foundational values. There is no evidence that the EU perceives itself as a follower or merely an observer in the global AI Ethics discourse. Based on the Double-Decisive test, hypothesis 1 (H1) is confirmed. Given the positive result of the Double-Decisive test, there is no need to perform any further tests to validate this hypothesis.

4.2 The US's approach to AI Ethics

This chapter marks the second stage of the empirical analysis. The objective is to outline the US's approach to AI Ethics over time and thereby provide a comprehensive empirical basis for later comparison. The first step is to identify the relevant documents, based on the given selection criteria (see Table 4). Next, the documents will be analyzed using the document analysis method of 'skimming,

reading, and interpretation' Bowen (2009). Chapter 3.4 outlines a coding scheme for the second stage, which this analysis will closely follow. This coding scheme differs from the EU's document analysis in that it includes additional categories, that are relevant for testing H2-H5. However, as with the EU document analysis, this chapter will also identify the US AI Ethics principles, concerns, suggested actions, and definitions, as well as how these evolved over time. Based on the coding scheme, relevant text passages will be identified and documented during the skimming and reading phase.

Table 4: Documents selected for the analysis of the US's approach to AI Ethics

No.	*Title*	*Date*
8	Executive Order 13859 Maintaining American Leadership in Artificial Intelligence (Federal Register 2019)	11.02.2019
9	A Plan for Federal Engagement in Developing Technical Standards and Related Tools (NIST 2019)	09.08.2019
10	Guidance for Regulation of Artificial Intelligence Applications (OMB 2020)	17.11.2020
11	Executive Order 13960 Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government (Federal Register 2020)	03.12.2020
12	National Artificial Intelligence Initiative Act (US Congress 2020)	03.12.2020
13	Draft Taxonomy of AI Risk (NIST 2021)	15.10.2021
14	Blueprint for an AI Bill of Rights (The White House 2022a)	04.10.2022
15	Artificial Intelligence Risk Management Framework (NIST 2023)	26.01.2023

Source: author's own.

Based on the findings, a comprehensive analysis of the US approach to AI Ethics over time will be conducted. The text passages, including citation number, citation, and category can be found in the Appendix[3] and will be referenced in square brackets. The analysis is presented chronologically.

Document analysis of the US's approach to AI Ethics

The first specifically AI related document within the scope of this research's analysis is 'Executive Order 13859: Maintaining American Leadership in Artificial Intelligence' (February 2019). The Executive Order (EO) was signed by then-President Donald J. Trump and aims to promote and maintain American leadership in AI research and development. While the order primarily directs the prioritization of investments in AI-related research and development as well as emphasizes the importance of preparing the American workforce for AI. It also stresses the importance of protecting American values in the development and application of AI [341-344]. Public trust is identified as a necessary precondition for maximizing the potential benefits of AI [342]. As a result, the National Institute of Standards and Technology (NIST) is directed to 'issue a plan for Federal engagement in the development of technical standards and related tools in support of reliable, robust, and trustworthy systems that use AI technologies.' [346]. Here, three AI-related principles - reliable, robust, trustworthy - are mentioned but not further specified. The priority is clearly enabling innovation and maintaining the 'economic and national security of the United States' [341].

[3] The passage numbers mentioned can be found in the appendix, which can be accessed via the following link: https://ib.uni-koeln.de/sites/jaeger/publikationen/aipa/8_Appendix_AIPA_12024.pdf.

Following the EO, NIST published 'A Plan for Federal Engagement in Developing Technical Standards and Related Tools' (August 2019). The plan identifies nine areas of focus for AI standards, including safety, risk management and trustworthiness [348]. The document identifies 'increasing trust in AI technologies as a key element in accelerating their adoption for economic growth and future innovations that can benefit society' [357], and highlights the importance of shaping international AI standards in a way that is favorable to the US [360]. According to the document, 'trustworthiness standards include guidance and requirements for accuracy, explainability, resiliency, safety, reliability, objectivity, and security' [349, 364]. These standards should be considered early in the design phase of AI development [357]. The term 'reliable, robust and trustworthy AI' – as mentioned in the EO – is very prominent throughout the document. One of the prioritized characteristics related to AI standards is 'human-centered to ensure that human interactions and values – including abilities, disabilities and diversity – are considered during AI data collection, model development, testing and deployment.' [375]. Sensitivity to ethical considerations is also highlighted as a priority characteristic, for 'identifying and minimizing bias, and incorporating provisions that protect privacy and reflect the broader community's notions of acceptability' [376]. According to the document, principles are the basis for standards. This requires a broad consensus on AI-related principles before common standards can be agreed upon. The document recognizes the importance of international organizations in this endeavor, noting that the US has adopted the OECD AI principles [368]. Standards should be developed carefully; according to the document 'the degree of potential risk presented by particular AI technologies will help to drive decision making about the need for specific AI standards and standards-related tools' [374].

Another result of EO 13859 was the memorandum 'Guidance for Regulation of Artificial Intelligence Applications' issued by the Director of the Office of

Management and Budget (OMB) in November 2020. This memorandum was intended to inform federal agencies on the development of regulatory and non-regulatory approaches to AI, as well as ways to reduce barriers to the development and adoption of AI. While it recognizes the potential need for 'a regulatory approach that fosters innovation, growth, and engenders trust, while protecting core American values' [379], it states that 'Federal agencies must avoid regulatory or non-regulatory actions that needlessly hamper AI innovation and growth' [382]. The document lays out 10 'Principles for the stewardship of AI Applications' [382]:

- Public Trust in AI: AI's potential must be balanced with the risk to privacy and civil liberty, and ensuring trustworthy applications for public trust [383].
- Public Participation: Agencies should promote public input in AI rulemaking and inform on AI standards, respecting legal limits [384].
- Scientific Integrity and Information Quality: AI approaches should emphasize scientific integrity, transparency, and reliable data training [385].
- Risk Assessment and Management: AI regulations should use risk-based evaluations, balancing harm, and benefits, without inhibiting innovation [386].
- Benefits and Costs: Agencies should weigh the societal costs and benefits of AI, considering its impact on existing systems and optimizing for net advantages [387].
- Flexibility: Agencies should use flexible, performance-based AI regulations, ensuring international competitiveness [388].
- Fairness and Non-Discrimination: Agencies should assess AI for potential biases, prioritizing fairness and examining discrimination impacts [389].
- Disclosure and Transparency: AI transparency can enhance trust; agencies should disclose AI use based on its impact and context [390].
- Safety and Security: Agencies should evaluate AI's potential for bias, ensuring regulations emphasize fairness [391].

- Interagency Coordination: Agencies should collaborate for a consistent approach to AI oversight, ensuring American innovation and values are upheld [392].

These principles emphasize a balanced AI approach, blending innovation with transparency, fairness, and a focus on public trust, while ensuring international competitiveness.

The EO 13960 'Promoting the Use of Trustworthy AI in the Federal Government', signed by President Donald J. Trump in December 2020, presented the first set of ethical principles published by the US. It set the goal of fostering the adoption and acceptance of AI and identified that this goal highly depends on public trust and confidence [398]. The design, development, acquisition, and use of AI in government should follow these principles [403-410]:

- Lawful and respectful, thereby aligning with the nation's values and legal frameworks.
- Purposeful and performance-driven, using AI where benefits outweigh manageable risks.
- Accurate, reliable, and effective, in line with the use case the AI was trained for.
- Safe, secure, and resilient, with regards to vulnerabilities, manipulation, and exploitation.
- Understandable, operations and outcomes being clear to relevant experts and users.
- Responsible and traceable, with clear human roles thorough documentation.
- Regularly monitored, with frequent testing of principles and deactivation mechanisms.
- Transparent, including disclosure to stakeholders, congress, and public, when legal.
- Accountable, enforcing AI safeguards, monitoring compliance, and training personnel.

The first piece of AI-related legislation from the US congress – The 'National AI Initiative Act' – was released in December 2020. In this legislation, the US Congress outlined its strategic vision for the US's approach to AI, with the underlying goal of maintaining the US's competitive edge in AI. While primarily calling for increased investment in AI research and development, promoting AI education, preparing the US's workforce for AI and interagency cooperation it also underscored the importance of addressing ethical, societal and safety implications of AI [414, 416, 420, 440]. Accordingly, the NIST was tasked with advancing AI standards [426] and developing a voluntary 'Risk Management Framework' (AI RMF) [435].

> The framework shall (1) identify and provide standards, guidelines, best practices, methodologies, procedures and processes for (A) developing trustworthy artificial intelligence systems; (B) assessing the trustworthiness of artificial intelligence systems; and (C) mitigating risks from artificial intelligence systems; (2) establish common definitions and characterizations for aspects of trustworthiness, including explainability, transparency, safety, privacy, security, robustness, fairness, bias, ethics, validation, verification, interpretability, and other properties related to artificial intelligence systems that are common across all sectors; (3) provide case studies of framework implementation; (4) align with international standards, as appropriate; (5) incorporate voluntary consensus standards and industry best practices; and (6) not prescribe or otherwise require the use of specific information or communications technology products or services [435-444].

In order to support the development of the AI RMF, NIST issued a formal request for information, which was answered with the white paper 'Taxonomy of AI Risk' (October 2021). The purpose of this white paper was to contextualize the characteristics of trustworthy AI as proposed by the NIST in its request for information. It distinguishes between three different categories of characteristics of trustworthy systems. Accuracy, reliability, robustness, and resilience, which have been identified as technical attributes [459-462]. Explainability, interpretability, safety, and managing bias have been identified as socio-technical attributes [464-469] as well as fairness, accountability, and transparency as guiding principles that contribute to AI trustworthiness [472-474]. The definitions provided can be found

in the appendix[4] of this paper, as it is beyond the scope of this analysis to provide them. The OECD AI principles, key principles from the EU's Ethics Guidelines for Trustworthy AI, and the EO 13960 principles were specifically mentioned as policy guidelines upon which the White Paper based its findings [452-454].

While the RMF was still being developed, the White House Office of Science and Technology Policy (OSTP) published 'The Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People' in October 2022. This blueprint includes five principles to help organizations guide the design, use, and deployment of automated systems while ensuring the protection of rights [480]. It also provides practical steps on how to potentially implement these principles. The five principles are:

- Safe and effective systems: You should be protected from unsafe or ineffective systems [484].
- Algorithmic Discrimination Protections: You should not face discrimination by algorithms, and systems should be used and designed in an equitable way [486].
- Data Privacy: You should be protected from abusive data practices via built-in protections and you should have agency over how data about you is used [491].
- Notice and Explanation: You should know that an automated system is being used and understand how and why it contributes to outcomes that impact you [495].
- Human Alternatives, Consideration, and Fallback: You should be able to opt out, where appropriate, and have access to a person who can quickly consider and remedy problems you encounter [499].

[4] The passage numbers mentioned can be found in the appendix, which can be accessed via the following link: https://ib.uni-koeln.de/sites/jaeger/publikationen/aipa/8_Appendix_AIPA_12024.pdf.

The most recent and – in terms of AI Ethics – most comprehensive document is the aforementioned results of the AI Initiative Act, the 'AI Risk Management Framework' (AI RMF), published by NIST in January 2023. The AI RMF aims to address the diverse risks associated with the deployment and use of AI systems, and provides detailed methods for framing AI-related risks and how they should be governed, mapped, measured, and managed [512]. It is intended to be flexible and non-binding. Existing regulations and guidelines should be prioritized over the AI RMF. The first key attribute of the AI RMF is that it should 'be risk-based, resource-efficient, pro-innovation, and voluntary' [548]. Which AI system is high risk is up to the organization applying the AI RMF. The AI RMF recommends that the highest risk AI systems should be prioritized. AI systems with unacceptable negative risk levels must be developed and deployed in a safe way until the risk is sufficiently manageable [521]. The document defines risk as 'the composite measure of an event's probability of occurring and the magnitude or degree of the consequences of the corresponding event' [514]. To determine risk levels, the AI RMF articulates characteristics of trustworthy AI systems that need to be balanced against the context in which the AI system is used:

- Valid and reliable: Validation ensures that AI systems meet specific requirements, with inadequate systems posing increased risks and decreasing trustworthiness [527]. Reliability in AI systems refers to their consistent performance as expected, under defined conditions, throughout their intended lifespan [528]. Accuracy in AI denotes how closely computational results match true values, and it's essential to test these results on representative data sets, with comprehensive documentation, including potential discrepancies between different data segments [530]. Robustness in AI refers to the system's capability to consistently perform well under varying conditions, ensuring functionality even in unanticipated scenarios and minimizing potential damage when operating in unexpected settings [531]. Accuracy and robustness are essential to the validity and trustworthiness of AI systems [529]. Assessing validity,

accuracy, robustness, and reliability enhances the AI's trustworthiness. Prioritizing risk management minimizes potential harm, and human intervention may be necessary for undetected or uncorrected AI errors [532].

- Safe: AI systems should ensure human and environmental safety through responsible practices, informed use, proactive risk management, early safety planning, and mechanisms for real-time adjustments [533, 534].
- Secure and resilient: AI systems should be resilient to unexpected events and secure against threats, relying on existing standards and considering both intended and unintended uses [535].
- Accountable and transparent: Accountability in AI is crucial and requires clear responsibility for AI outcomes in different contexts, especially when the consequences are severe [539]. Transparency in AI means providing accessible information about the system's design, training, and functionality to promote understanding and trust, while considering the balance with proprietary information [537].
- Explainable and interpretable: Explainability describes the inner workings of AI systems, while interpretability refers to the significance of their outputs; both increase understanding and trust in the system's functionality [541].
- Privacy-enhanced: Privacy protects human autonomy and dignity, guiding the AI system design with values such as anonymity and control, while balancing trade-offs with security, bias, and transparency, as AI may pose new risks to individual identification and information disclosure [542, 543].
- Fair with harmful bias managed: Fairness in AI seeks to address harmful biases across systemic, computational, and human-cognitive categories, yet mitigating these biases doesn't guarantee fairness; unchecked, AI can amplify biases, impacting transparency and societal equity [544-547].

According to the AI RMF, these characteristics influence each other, and therefore risk management requires balancing trade-offs between them [526].

The US approach to AI Ethics has progressed from an initial focus on technological leadership and economic considerations to a more balanced approach that incorporates ethical principles, transparency, and accountability. The US has gradually acknowledged the importance of public trust and the need to address ethical and societal implications of AI, aligning with international standards and principles. This evolution demonstrates an increasing awareness of the multifaceted challenges and opportunities presented by AI.

4.3 Diffusion mechanisms of the EU's AI Ethics approach to the US

Stage three of the empirical analysis has two objectives. The first objective is to identify *if* the US's approach to AI Ethics has shifted towards the EU's approach. To achieve this, the same hoop test was chosen for each hypothesis: the US changes its AI Ethics approach and moves towards the EU's approach. This provides a necessary condition for further inquiry. If the evidence does not pass the test, the hypothesis (and therefore all hypotheses) can be refuted.

For a close comparison of the EU's and US's approaches to AI Ethics over time, a detailed timeline was created as proposed in the chapter about process tracing. Figure 1 provides an overview of the sequence of events for both the EU and the US based on the previous document analysis. The EU had already taken initial steps in terms of AI strategy and respective AI Ethics considerations before the US published its first document (as considered in this analysis). The communication 'AI for Europe' (document 1) highlights the importance of developing and deploying AI based on European values and fundamental rights due to the associated risks. The document introduces several principles, including explainability, human-centric, safety, security, inclusion, transparency, non-discrimination, privacy, dignity, and fairness, without providing clear definitions.

Table 5: Timeline of EU and US AI Ethics related documents

AI for Europe (1) *04/18* **Introduced:** values, fundamental rights, trust, explainability, human-centric, safety, security, inclusion, transparency, non-discrimination, privacy, dignity, fairness	**Coordinated Plan on AI 2018 (2)** *12/18* **Introduced:** Trustworthy AI **Defined:** Trust: Predictable; responsible; verifiable; respecting fundamental rights; following ethical rules	**Ethics Guidelines for Trustworthy AI (3)** *04/19* **Defined:** Trustworthy AI: • Lawful, ethical (respectful), robust (safe, secure, reliable) (components) • Respect for human autonomy, prevention of harm, fairness, explicability (principles) • Human agency and oversight; technical robustness (resilience and security, accuracy, reliability) and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing (sustainability); accountability (responsibility) (requirements) Human centric Robust	**P&I Rec. for trustworthy AI (4)** *07/19* **Introduced:** Risk-based approach to trustworthy AI **Defined:** Risk: Adverse impacts of all kinds, both individual and societal.	**White Paper on AI (5)** *02/20* **Introduced and defined:** High-risk AI application: • Significant risks can be expected to occur in specific sector and significant risks are likely to arise by using it (criteria). • Training data; data and record-keeping; information to be provided; robustness and accuracy; human oversight; specific requirements for certain particular AI applications, e.g. biometric identification (requirements)	**AI Act (7)** *04/21* **Defined:** Risk-based approach: • Minimal, limited, high, unacceptable risk (risk levels) • Unacceptable: contravene EU's values by violating fundamental rights. • High: high quality data; documentation and traceability; transparency; human oversight; accuracy and robustness		
EO 13859 (8) *02/19* **Introduced:** *Trust,* reliable, *robust, trustworthy*	**NIST Plan (9)** *08/19* **Introduced:** *Safety, risk management, trustworthiness* (3/9 standards); *Human-centered, bias reduction, privacy*, acceptability, *risk-based approach* **Defined:** *Trustworthiness*: *Accuracy, explainability, resilience, safety, reliability*, objectivity, *security* (guidance & requirements)	**Guidance for Regulation (10)** *11/20* **Introduced:** 10 Principles for Stewardship of AI applications (partially AI ethics related): Public trust; scientific integrity and information quality; risk assessment and management; *fairness* and *non-discrimination*; disclosure and *transparency*; safety and security	**EO 13960 (11)** *12/20* **Defined**: *Trustworthy AI:* *Lawful* and *respectful*; purposeful and performance-driven; *accurate, reliable,* and effective; *safe, secure,* and *resilient*; *understandable*; *responsible* and traceable; regularly monitored; *transparent*; *accountable* (principles)	**National AI Initiative Act (12)** *12/20* Tasked NIST with AI RMF: "identify best practices", "establish common definitions", "align with international standards"	**Taxonomy of AI Risk (13)** *10/21* **Defined:** *Trustworthy AI*: *Accuracy*; *reliability*; *robustness*; *resilience*; *explainability*; interpretability; *safety* and *managing bias*; *fairness, accountability, transparency*	**Blueprint for an AI Bill of Rights (14)** *10/22* **Introduced:** *Safe* and effective systems; *algorithmic discrimination protections*; *data privacy*; notice and *explanation*; human alternatives, consideration, and *fallback* (principles)	**AI RMF (15)** *01/23* **Introduced:** **High-risk AI application** **Defined:** *Risk* *Trustworthy AI:* Valid and *reliable*; *safe*; *secure and resilient*; *accountable* and *transparent*; *explainable* and *interpretable*; *privacy-enhanced*; *fair with harmful bias managed* (characteristics)

Source: author´s own.

The key ethical principle that emerged for the first time was *trust*. This led to the significant term 'Trustworthy AI' being introduced in the Coordinated Plan 2018 (2). This document defined key terms such as trust, predictable, responsible, verifiable, respecting fundamental rights, and adherence to ethical rules. EO 13859 (8) strongly published shortly after document 2 and, while not emphasizing ethical considered trust, reliability, robustness, and trustworthiness. The importance of trust was emphasized, similar to the EU's previous communications. Unfortunately, there is little convergence observable at this point. The next pivotal EU document is the Ethics Guidelines (3), which were published in April 2019. The document defines 'Trustworthy AI' into three components, four principles, and seven requirements (see Table 5). Additionally, the terms 'robust' and 'human-centric' were defined, highlighting the importance of human-centricity in the EU's understanding of AI Ethics. In July 2019, the Policy and Investment Recommendations (4) importantly introduced the risk-based approach to trustworthy AI and defined risk. This understanding of trustworthy AI and the risk-based approach remains central to the EU's approach to AI. The NIST Plan (9) followed shortly afterwards and, while not as extensive as the EU's documents, introduced many new AI Ethics related principles and considerations for the US. Most importantly, the EU's understanding of AI Ethics has evolved with the introduction of a human-centered approach, a recommendation for a risk-based approach, and a definition of trustworthiness. These principles are largely consistent with previous EU documents, indicating a significant step forward from EO 13859. It is important to note, however, that documents 3,4 and 9 all are recommendatory documents. When examining their further evolution, it becomes clear that they influenced the approach of both entities.

The White Paper on AI (5) from February 2020, in line with the afore proposed risk-based approach, introduced the category of 'High-risk AI application' and outlined proposed respective criteria and requirements. While the

Guidance for Regulation of AI Systems (10) from November 2020 aimed to reduce barriers to the development and adoption of AI and foster innovation. It also introduced the first clear set of AI-related principles, some of which were partially related to AI Ethics. One month after its introduction, the EO 13960 established the first set of ethical principles in the US. Although not entirely congruent with those promoted by the EU, there are considerable overlaps between the two (see figure 1). The first principle of this set is 'lawful and respectful', similar to the EU's three components of lawful, ethical (including respectful), and robust. This marks yet another shift from the US towards the EU's AI Ethics approach.

The AI Act (6) proposed a regulatory framework for AI-systems, with a focus on the High-risk AI systems. The most relevant introduction in terms of AI Ethics is the distinction between risk categories: minimal, limited, high, and unacceptable. The latest document published by the US, the AI RMF from January 2023 (15), similarly introduces the term 'high-risk AI application'. Moreover, Trustworthy AI characteristics are presented All of the terms, except for 'valid', can be found in the EU's definition of Trustworthy AI.

This high-level comparison demonstrates that, over time, the US has followed the EU's approach to AI step-by-step. While EO 13859 did not prioritize ethical considerations, the latest AI RMF and the Blueprint of AI Bill of Rights clearly prioritize ethical considerations. Over time, the US's understanding of AI Ethics has shifted significantly closer to the EU's understanding. The organization published several principles that align with the EU's ethical framework, emphasizing trustworthy AI and introducing a risk-based approach to AI systems. It is important to note that, to date, the US has not produced a legislative piece similar to the AI Act. All documents are recommendations, guidance, and non-binding. Nevertheless, there is a clear trend towards greater concern for AI Ethics and a shift towards the EU's approach to AI Ethics. As a result, the hoop test is passed. The rest

of the analysis will focus on presenting evidence of potential causal diffusion mechanisms between the EU and the US for each hypothesis.

H2: The US adopts a similar approach to AI Ethics due to coercive pressure from the EU.

- Smoking-Gun Test: The USA changes its AI Ethics approach, moving to the EU's approach, shortly after coercive pressure from the EU.
- Doubly Decisive Test: Direct evidence from the US that confirms the alignment with the EU's AI Ethics approach in response to coercive pressure from the EU.

No evidence was found to support for any form of coercive pressure from the EU, thus refuting this hypothesis. On the contrary, evidence of AI-related cooperation was found. Firstly, both the EU and the US signed the OECD's 'Recommendation of the Council on Artificial Intelligence' in May 2019, which outlined five 'Principles for responsible stewardship of trustworthy AI' [223, 368]. Secondly, the EU and the US established the EU-US Trade and Technology (TTC) council in 2021 [243].

Based on the assumption of the Life Cycle of Norms as described earlier, norm entrepreneurs use international organizations to promote their norms. The OECD principles and the TTC engagement may provide additional evidence for diffusion mechanisms. Therefore, a document analysis will be performed of a) the OECD recommendation document and b) joint statements and documents found in the context of the TTC. As shown in Table 6, the following documents were identified.

Table 6: Selected OECD and TTC documents

No.	*Title*	*Date*
16	OECD Recommendation of the Council on Artificial Intelligence (OECD 2023)	22.05.2019
17	EU-US Inaugural Joint Statement of the TTC (The White House 2021)	29.09.2021
18	TTC Joint Roadmap on Evaluation and Measurement Tools for Trustworthy AI and Risk Management (European Commission 2023c)	16.05.2022
19	EU-US 2nd Joint Statement of the TTC (The White House 2022b)	01.12.2022
20	EU-US 3rd Joint Statement of the TTC (The White House 2022c)	05.12.2022
21	EU-US 4th Joint Statement of the TTC (The White House 2023)	31.05.2023
22	EU-US Terminology and Taxonomy for AI (European Commission 2023b)	23.05.2023

Source: author's own.

The following coding scheme was applied during the document analysis (slightly adapted from the coding scheme of the US document analysis).

Table 7: Codes and categories for the analysis of the OECD principles and TTC documents

Code name	*Category*
B1	Any ethical principle/value that is mentioned.
B2	Any AI-related risk/challenge that is mentioned
B3	Recommended AI Ethics related action.
B4	AI Ethics related definition.
B5	Mention of international cooperation.
B6	Any mention that highlights competitive pressure.

Source: author's own.

H3: The US adopts a similar approach to AI Ethics based on learning from the EU's experiences and frameworks.

Making a case for a diffusion mechanism of learning is challenging. Learning is outcome-oriented, and AI Ethics, especially its governance, is still in its early stages. The evidence tests require either a reference to the EU's *successful* approach to AI Ethics (Smoking-gun) or a direct statement confirming the adoption of a similar approach based on learning from the EU's experience (doubly decisive). An argument could be made that the EU's approach to AI Ethics is internationally recognized and perceived as a benchmark (Jobin et al. 2019; Floridi 2021c, 2021a). As a result, their approach could be perceived as 'successful' by the US. Concrete evidence for this scenario cannot be found in the US's documents. The EU's AI Ethics Guidelines are referred to once [453] as one of three referenced documents for the formulation of principles. In the same paragraph, the text references both the OECD AI principles and the US principles. The OECD AI principles were adopted by the OECD member states in May 2019, one month after the publication of the EU's AI Ethics Guidelines. The inherent principles for a 'responsible stewardship of trustworthy AI' are a) inclusive growth, sustainable development, and well-being, b) human-centred values and fairness, c) transparency and explainability, d) robustness security and safety as well as e) accountability [550]. Each of these principles aligns with the principles outlined in the EU's Ethics Guidelines, except for 'inclusive growth'. This indicates that the EU had substantial influence on the formulation of the OECD principles, as stated in the EU'S AI White Paper [223]. Based on this, one *could* argue for a potential diffusion mechanism of learning as the US refers to EU principles and potentially EU-influenced principles. However, the evidence is too weak for this thesis as learning as a diffusion mechanism is based on the perception of a successful outcome due to a policy. Evaluating the success of the EU's AI Ethics approach is beyond the scope of this thesis; therefore, the hypothesis is rejected.

H4: The US emulates a similar approach to AI Ethics as the EU's due to a perception of appropriateness.

Emulation is a normative diffusion mechanism that could potentially serve as a diffusion mechanism for this research. In the case of emulation, political actors make decisions based on 'appropriateness' rather than 'consequences' (Checkel 2005; Gilardi 2013). Norm entrepreneurs also push their norms into the international system, which has been identified as the self-perceived role of the EU. Therefore, the evidence suggests that the US needs to adjust its approach to AI Ethics to align more closely with the EU's approach out of appropriateness. The question is: Does the US adopt these norms due to appropriateness or due to consequences? The answer could potentially be both. However, the US documents are increasingly highlighting the risks associated with AI Ethics and formulating specific AI ethic principles accordingly. From the beginning it highlighted the importance of adherence to human rights and values. This can be understood as an approach based on appropriateness. Moreover, in the second joint TTC statement, the EU and US highlight the importance of international cooperation to oppose rights-violating AI systems such as social scoring [581]. While other 'consequence'-oriented motivations may also be present (as will be outlined in H5), the documents from the US consistently convey a sense of ethical appropriateness. This thesis argues that both perspectives can coexist, representing different opinions within a democratic system. Therefore, the hypothesis passes the smoking gun test. No direct statement that confirms the alignment with the EU's AI Ethics approach out of appropriateness could be found, therefore the doubly decisive test is not passed. As the smoking gun test is sufficient to confirm causality, the H4 is valid.

H5: The US adopts a similar approach to AI Ethics to the EU's in response to competitive pressure.

As opposed to emulation, competition as a diffusion mechanism is very outcome-oriented. The ultimate goal is economic advantage. This thesis's analysis strongly

supports this diffusion mechanism. To test this hypothesis, it is necessary to recognize the economic and competitive benefits of promoting ethical AI principles. Although the approach to AI Ethics in the US has evolved over time and become more aligned with that of the EU, one aspect has remained consistent: the emphasis on the importance of spearheading the developments around AI globally. EO 13859 prioritized innovation and safeguarding the 'economic and national security of the United States' [341]. Trust was identified as an essential component in order in maximizing the potential benefits of AI [342]. The NIST plan suggests to 'strategically engage with international parties to advance AI standards for US economic and national security needs' [356] and highlights the importance of shaping international AI standards favorably for the US [360]. It acknowledges that 'lack of US stakeholder engagement in the development of AI standards can degrade the innovativeness and competitiveness of the US in the long term' [377]. Again, trust is identified as a 'key element in accelerating their adoption for growth and future innovations that can benefit society' [357]. The importance of creating public trust is emphasized in EO 13960 [398, 401], which is echoed in the TTCs 3rd and 4th statements [596, 605]. The smoking gun test is passed as the EU emphasizes trustworthy AI and the US aligns ethical principles and standards over time (as presented under H1), with 'trust' being perceived as a necessary component for maximal economic advantage of AI. Although various text passages confirm an international alignment on AI standards for competitive benefits, none directly confirm the adoption of a similar AI Ethics approach by the US to the EU's. Therefore, the doubly decisive test is rejected. Nevertheless, as outlined earlier, passing the smoking-gun test is enough for causal inference. Due to the amount of evidence supporting the notion that competitive pressures influence the US's alignment with the EU's AI Ethics approach and the passing of the smoking-gun test, causality can be inferred. Thus, H5 – the US adopts a similar approach to the EU's in response to competitive pressure – is valid.

So, what is the effect of the EU's AI approach on the US? First of all, as the empirical analysis of the EU's approach clearly highlighted, the EU perceives itself as a global norm entrepreneur in the field of AI Ethics. The EU aims to set an example for other countries and seeks to influence other states through bilateral efforts and international platforms such as the OECD. Therefore, H1 is confirmed. While this finding was predictable, it was still crucial for the following empirical analysis of diffusion mechanisms. It is important to note that having a self-perceived responsibility and role as a norm entrepreneur in the realm of AI Ethics does not necessarily result in the actual influence of other countries. This is where potential diffusion mechanisms come into play – offering a lens through which to potentially identify causal mechanisms of policy diffusion between countries. The four diffusion mechanisms identified during the research design were coercion, learning, emulation, and competition, represented by H2-H5. To infer causality, process tracing and evidence tests were applied. The first step was to establish the necessary condition for the hypothesis, which was the hoop test. The analysis revealed that for each hypothesis, the US shifted its AI Ethics approach towards that of the EU over time. The US continuously emphasized ethical considerations around AI and gradually adopted multiple facets of the EU's AI Ethics approach. Especially the increasing centrality of trustworthy AI and the introduction of a risk-based approach to AI Ethics are a significant convergence. This convergence of the AI Ethics approach does not infer causality, yet. Further tests were conducted in order to infer causal diffusion mechanisms. The document analysis revealed no evidence of coercive pressure, thus H2 was rejected. The analysis indicated potential learning mechanisms based on certain indicators. References were made to the EU's Ethics Guidelines and the OECD principles, which closely align with the EU's principles for trustworthy AI. The analysis concluded that the evidence was not strong enough. According to the diffusion mechanism of learning, a *successful* policy outcome must be perceived. Evaluating the success of the EU'S AI Ethics approach and the respective perception was outside of the scope of this research, though.

Therefore, H3 was declined. In the next step, the analysis provides an argument for the existence of both a sense of appropriateness – and hence, an existing emulation diffusion mechanism – and a strong economically incentivized outcome-oriented competition diffusion mechanism. While the US continuously emphasized the importance of adherence to human rights and values, which indicates a sense of appropriateness and therefore emulation, a strong focus on the economic benefits – competition - of promoting ethical AI principles was also identified. Trust is identified as a key component for the public acceptance of AI technology, and as a result, the US is committed to fostering trustworthy AI. This thesis argues that emulation and competition are not contradictory. Therefore, both H4 and H5 were accepted based on passing their respective smoking-gun evidence tests.

5 Discussion

How can these findings be interpreted? As previously mentioned, this thesis has outlined limitations of the research designs, including the sole reliance on documents. Due to the scope of this research, it is impossible to factor in underlying mechanisms and motivations that are not depicted in the analyzed documents. For instance, learning could actually be an underlying diffusion mechanism, as was indicated, but the documents were unable to prove it. Likewise, it could be that the role of emulation was overemphasized, and the US represents a sense of appropriateness solely based on economic motivations. In this context, it is important to highlight a potential interpretative bias the author could have, as is often the case in qualitative research. Moreover, while the focus on the EU and the US does allow for a zoomed in analysis, it potentially neglects influences of other entities in the field of AI Ethics. If the EU's approach to AI Ethics were to be completely influenced by any other country, the design and scope of this study

would not be able to capture it. Therefore, arguing for diffusion mechanisms between the EU and the US would become more complicated.

Nevertheless, this thesis provides both empirical and theoretical value. First, it contributes to existing literature by uncovering potential diffusion mechanisms between the EU and the US, two major players in the field of AI. This is valuable as it enhances our understanding of the complex AI Ethics landscape. The empirical value is founded in the detailed chronological analysis of both the EU's and the US's evolutionary approach to AI Ethics over time. This can serve as a basis for further analysis, such as comparing it to other countries. The research employs theory-testing and process tracing, offering valuable theoretical insights. By testing H1, it highlights the relevance of underlying theories and concepts, such as Normative Power Europe and the Life Cycle of Norms. The combination of evidence tests, and diffusion mechanisms proved to be a fruitful combination for causal inference, which is often a problem in diffusion research. Moreover, it was highlighted that multiple diffusion mechanisms can be at play simultaneously. This could potentially be the subject of further research – investigating how and why different diffusion mechanisms overlap.

What is the outlook? As of now, neither the EU nor the US have a legally binding regulatory framework for AI Ethics. It is important to note that the European Parliament's negotiating position on the AI Act was adopted in June 2023. The next step is negotiating with EU countries in the Council about the final form of the AI Act (EU 2023). The law is expected to be passed by the end of 2023 or the beginning of 2024 (Sharp 2023). While studies have already projected the potential impact of the AI Act (Greenleaf 2021; Feldstein 2023; Birchfield et al. 2022), it will be crucial to reevaluate the situation once the AI Act is legally binding. Currently, the US has merely moved towards the EU's approach to ethics by publishing standards and guidelines for organizations to *voluntarily* implement. However, this thesis highlights the risks of ethics shirking and ethics bluewashing earlier and the need

to move from principles to practical implementation. The US has yet to take this step. The Brussels effect demonstrates that the GDPR, implemented by the EU became a global standard due to the EU's regulatory influence (Bradford 2020). It is uncertain whether the AI Act will follow the same approach, and whether the US will adopt a similar legally binding regulatory framework for AI that prioritizes ethics.

6 Conclusion

The central research question of this thesis was 'What is the influence of the EU's AI Ethics approach on the US?'. The research question was approached from two angles. Firstly, the EU and US approaches to AI Ethics over time are examined through a systematic document analysis and subsequent comparison, based on assumptions derived from Constructivism and Normative Power Europe. The analysis revealed that the EU considers itself a norm entrepreneur in the field of AI Ethics and aims to shape international norms accordingly. Additionally, it was found that the US has increasingly adopted the EU's approach to AI Ethics over time. For instance, in the Ethics Guidelines for Trustworthy AI published by the AI HLEG in 2019, the EU provided a precise outline of trustworthy AI and the relevant ethical considerations at an early stage. Although the US occasionally mentioned ethical considerations regarding AI, there was a significant increase in alignment with the EU's understanding of central terms such as 'trustworthy' AI. The second part of the empirical analysis sought to identify *why* the US's AI Ethics approach shifted towards the EU's approach. Based on the theoretical assumptions of diffusion theory, the analysis tested for four potential causal diffusion mechanisms: coercion, learning, emulation, and competition. In accordance with process tracing, evidence tests were applied in order to identify potential causal mechanisms between the respective sequences. No evidence was found for a coercive diffusion

mechanism. While some possible indications were identified for an underlying learning mechanism (due to references to EU policies in US documents), this thesis rejected the respective hypothesis. Learning as a diffusion mechanism is based on the perception of a successful policy of another country/organization. Within the research scope of this thesis, sufficient evidence to support this requirement was not found. Respective US documents consistently emphasized the importance of adhering to human rights and values when it comes to emulation and ethical considerations around AI. Therefore, the respective hypothesis was accepted, and emulation was identified as the underlying diffusion mechanism. Even stronger evidence was found for the diffusion mechanism of competition. Early on, public trust was identified as a crucial factor in gaining public acceptance of AI technology. Additionally, there was a strong interest in promoting AI-related economic growth and innovation. Therefore, it is logical for the US to shift towards the EU's emphasis on 'trustworthy AI' to maximize economic benefits.

In summary, this research has observed a substantial influence of the EU's approach on the US. The US's ethical understanding of AI is now much closer to that of the EU's than it was in 2018. Nevertheless, the US has not yet followed the EU in presenting a legally binding regulatory framework for AI and is still relying on guidelines and principles. These findings of this research provide both empirical and theoretical value to existing research. First, the in-depth analysis compares the EU and US approaches to AI Ethics from 2018 to 2023, providing a strong empirical basis that fills an existing research gap and provides potential paths for further research. The findings offer insights into the AI-related priorities of both regions and could be compared to those other countries or regions to identify potential causal relationships. The theoretical value of this study lies in emphasizing the relevance of the underlying theories and concepts, namely Normative Power Europe and Diffusion. The EU continues to present itself as a normatively influential actor within the international system. Moreover, combining process-tracing based evidence tests and diffusion theory based causal mechanisms proved to be well

equipped at identifying underlying causal diffusion mechanisms, which otherwise can often be a challenge in diffusion research.

This research provides a temporal snapshot of the highly dynamic field of AI and AI Ethics. It is important to continuously conduct research in this field. In light of the findings of this thesis, it will be necessary to continuously scrutinize the approach of both the EU and the US to AI Ethics to fully understand their respective underlying motivations – especially with the expected adoption of the EU AI Act at the end of 2023 or the beginning of 2024.

7 References

AI HLEG (2019a): Ethics Guidelines for Trustworthy AI. With assistance of AI HLEG. Edited by AI HLEG. Available online at https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, updated on 10/21/2023, checked on 10/21/2023.

AI HLEG (2019b): Policy and investment recommendations for trustworthy Artificial Intelligence. Available online at https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence, updated on 10/21/2023, checked on 10/21/2023.

Beach, Derek; Pedersen, Rasmus (2019): Process-Tracing Methods. Ann Arbor, MI: University of Michigan Press.

Biersteker, Thomas J. (1990): Reducing the Role of the State in the Economy: A Conceptual Exploration of IMF and World Bank Prescriptions. In *Int Stud Q* 34 (4), p. 477. DOI: 10.2307/2600608.

Birchfield, Vicki L.; Roy, Varun; Sreedhar, Vignesh (2022): The EU's potential to lead in "ethical and secure" artificial intelligence: last, best hope? In *J Transatl Stud* 20 (3-4), pp. 299–327. DOI: 10.1057/s42738-023-00101-3.

Boddington, P. (2017): Towards a code of ethics for artificial intelligence: Springer. Available online at https://link.springer.com/content/pdf/10.1007/978-3-319-60648-4.pdf.

Bostrom, Nick; Yudkowsky, Eliezer (2018): The Ethics of Artificial Intelligence. In : Artificial Intelligence Safety and Security: Chapman and Hall/CRC, pp. 57–69. Available online at https://www.taylorfrancis.com/chapters/edit/10.1201/9781351251389-

4/ethics-artificial-intelligence-nick-bostrom-eliezer-yudkowsky?context=ubx.

Bowen, Glenn A. (2009): Document Analysis as a Qualitative Research Method. In *Qualitative Research Journal* 9 (2), pp. 27–40. DOI: 10.3316/qrj0902027.

Bradford, Anu (2012): The Brussels Effect. In *Nw. U. L. Rev.* 107, p. 1. Available online at https://heinonline.org/HOL/Page?handle=hein.journals/illlr107&id=1&div=3&collection=journals.

Bradford, Anu (2020): The Brussels effect. How the European Union rules the world. New York, NY: Oxford University Press.

Brooks, Sarah M. (2005): Interdependent and Domestic Foundations of Policy Change: The Diffusion of Pension Privatization Around the World. In *Int Stud Q* 49 (2), pp. 273–294. DOI: 10.1111/j.0020-8833.2005.00345.x.

Brundage, Miles; Avin, Shahar; Clark, Jack; Toner, Helen; Eckersley, Peter; Garfinkel, Ben et al. (2018): Malicious Use of Artificial Intelligence : Forecasting, Prevention, Mitigation. Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Security, Electronic Frontier Foundation, OpenAI.

Butcher, James; Beridze, Irakli (2019): What is the State of Artificial Intelligence Governance Globally? In *The RUSI Journal* 164 (5-6), pp. 88–96. DOI: 10.1080/03071847.2019.1694260.

Cao, Xun; Prakash, Aseem (2012): Trade Competition and Environmental Regulations: Domestic Political Constraints and Issue Visibility. In *The Journal of Politics* 74 (1), pp. 66–82. DOI: 10.1017/S0022381611001228.

Checkel, Jeffrey T. (2005): International Institutions and Socialization in Europe: Introduction and Framework. In *Int Org* 59 (04). DOI: 10.1017/s0020818305050289.

Collier, David (2011): Understanding Process Tracing. In *APSC* 44 (04), pp. 823–830. DOI: 10.1017/s1049096511001429.

Dafoe, Allan (2018): AI governance: a research agenda. Available online at https://www.fhi.ox.ac.uk/wp-content/uploads/govai-agenda.pdf.

Daly, Angela; Hagendorff, Thilo; Li, Hui; Mann, Monique; Marda, Vidushi; Wagner, Ben et al. (2019): Artificial Intelligence, Governance and Ethics: Global Perspectives.

Damro, Chad (2012): Market power Europe. In *Journal of European Public Policy* 19 (5), pp. 682–699. DOI: 10.1080/13501763.2011.646779.

Dignum, Virginia (2018): Ethics in artificial intelligence: introduction to the special issue. In *Ethics Inf Technol* 20 (1), pp. 1–3. DOI: 10.1007/s10676-018-9450-z.

Djeffal, Christian; Siewert, Markus B.; Wurster, Stefan (2022): Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strategies. In *Journal of European Public Policy* 29 (11), pp. 1799–1821. DOI: 10.1080/13501763.2022.2094987.

Elkins, Zachary; Guzman, Andrew T.; Simmons, Beth A. (2006): Competing for Capital: The Diffusion of Bilateral Investment Treaties, 1960–2000. In *Int Org* 60 (04). DOI: 10.1017/s0020818306060279.

EU (2023): EU AI Act: first regulation on artificial intelligence | News | European Parliament. Available online at https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, checked on 10/21/2023.

European Commission (Ed.) (2018a): Artificial Intelligence for Europe. European Commission. Available online at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN, updated on 10/21/2023, checked on 10/21/2023.

European Commission (Ed.) (2018b): Coordinated Plan on Artificial Intelligence. With assistance of European Commission. Available online at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM:2018:795:FIN, updated on 10/21/2023, checked on 10/21/2023.

European Commission (2020): White Paper on Artificial Intelligence: a European approach to excellence and trust. With assistance of European Commission. Edited by European Commission. Available online at https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en, updated on 2/19/2020, checked on 10/21/2023.

European Commission (2021): LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. With assistance of European Commission. Edited by European Commission. Available online at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, updated on 10/21/2023, checked on 10/21/2023.

European Commission (Ed.) (2023a): Coordinated Plan on Artificial Intelligence 2021 Review. With assistance of European Commission. Available online at https://digital-strategy.ec.europa.eu/en/library/coordinated-plan-artificial-intelligence-2021-review, updated on 10/20/2023, checked on 10/21/2023.

European Commission (Ed.) (2023b): EU-US Terminology and Taxonomy for Artificial Intelligence. With assistance of European Commission.

Available online at https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence, updated on 10/20/2023, checked on 10/21/2023.

European Commission (Ed.) (2023c): TTC Joint Roadmap for Trustworthy AI and Risk Management. With assistance of European Commission. Available online at https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management, updated on 10/20/2023, checked on 10/21/2023.

Federal Register (2019): Maintaining American Leadership in Artificial Intelligence. Available online at https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence, updated on 10/21/2023, checked on 10/21/2023.

Federal Register (2020): Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. Available online at https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government, updated on 10/21/2023, checked on 10/21/2023.

Feldstein, Steven (2023): Evaluating Europe's push to enact AI regulations: how will this influence global norms? In *Democratization*, pp. 1–18. DOI: 10.1080/13510347.2023.2196068.

Finnemore, Martha; Hollis, Duncan B. (2016): Constructing Norms for Global Cybersecurity. In *Am. j. int. law* 110 (3), pp. 425–479. DOI: 10.1017/s0002930000016894.

Finnemore, Martha; Sikkink, Kathryn (1998): International Norm Dynamics and Political Change. In *Int Org* 52 (4), pp. 887–917. DOI: 10.1162/002081898550789.

Fjeld, Jessica; Achten, Nele; Hilligoss, Hannah; Nagy, Adam; Srikumar, Madhulika (2020): Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI.

Floridi, Luciano (2021a): Establishing the Rules for Building Trustworthy AI. In Luciano Floridi (Ed.): Ethics, Governance, and Policies in Artificial Intelligence, vol. 144. Cham: Springer International Publishing (Philosophical Studies Series), pp. 41–45.

Floridi, Luciano (2021b): Introduction – The Importance of an Ethics-First Approach to the Development of AI. In : Ethics, Governance, and Policies in Artificial Intelligence: Springer, Cham, pp. 1–4. Available online at https://link.springer.com/chapter/10.1007/978-3-030-81907-1_1.

Floridi, Luciano (2021c): Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. In : Ethics, Governance, and Policies in Artificial Intelligence: Springer, Cham, pp. 81–90. Available online at https://link.springer.com/chapter/10.1007/978-3-030-81907-1_6.

Floridi, Luciano; Cowls, Josh; Beltrametti, Monica; Chatila, Raja; Chazerand, Patrice; Dignum, Virginia et al. (2018): AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. In *Minds and machines* 28 (4), pp. 689–707. DOI: 10.1007/s11023-018-9482-5.

Future of Life Institute (2023): Pause Giant AI Experiments: An Open Letter - Future of Life Institute. Available online at https://futureoflife.org/open-letter/pause-giant-ai-experiments/, updated on 9/18/2023, checked on 10/22/2023.

Garcia, Eugenio V. (2022): Multilateralism and Artificial Intelligence : What Role for the United Nations? In Maurizio Tinnirello (Ed.): The global politics of artificial intelligence. First edition. Boca Raton, FL, London, New York:

CRC Press Taylor & Francis Group (Chapman & Hall/CRC artificial intelligence and robotics series), pp. 57–84. Available online at https://www.taylorfrancis.com/chapters/edit/10.1201/9780429446726-3/multilateralism-artificial-intelligence-eugenio-garcia.

Gerring, John (2005): Causation. In *Journal of Theoretical Politics* 17 (2), pp. 163–198. DOI: 10.1177/0951629805050859.

Gibbs, Samuel (2014): Elon Musk: artificial intelligence is our biggest existential threat. In *The Guardian*, 10/27/2014. Available online at https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat, checked on 10/22/2023.

Gilardi, Fabrizio (2010): Who Learns from What in Policy Diffusion Processes? In *American Journal of Political Science* 54 (3), pp. 650–666. DOI: 10.1111/j.1540-5907.2010.00452.x.

Gilardi, Fabrizio (2013): Transnational Diffusion: Norms, Ideas, and Policies. In : Handbook of International Relations. 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom: SAGE Publications Ltd, pp. 453–477.

Greenleaf, Graham (2021): The 'Brussels Effect' of the EU's 'AI Act' on Data Privacy Outside Europe. In *Privacy Laws & Business International Report*.

Hagendorff, Thilo (2022): Blind spots in AI ethics. In *AI Ethics* 2 (4), pp. 851–867. DOI: 10.1007/s43681-021-00122-8.

Hagendorff, Thilo (2023): AI ethics and its pitfalls: not living up to its own standards? In *AI Ethics* 3 (1), pp. 329–336. DOI: 10.1007/s43681-022-00173-5.

Hedström, Peter; Ylikoski, Petri (2010): Causal Mechanisms in the Social Sciences. In *Annu. Rev. Sociol.* 36 (1), pp. 49–67. DOI: 10.1146/annurev.soc.012809.102632.

Hern, Alex (2016): Stephen Hawking: AI will be 'either best or worst thing' for humanity. In *The Guardian*, 10/19/2016. Available online at https://www.theguardian.com/science/2016/oct/19/stephen-hawking-ai-best-or-worst-thing-for-humanity-cambridge, checked on 10/22/2023.

Horowitz, Michael C. (2018): Artificial Intelligence, International Competition, and the Balance of Power. In *Texas National Security Review*, 2018. Available online at https://tnsr.org/2018/05/artificial-intelligence-international-competition-and-the-balance-of-power/, checked on 6/21/2020.

Jahn, Detlef (2022): Diffusion. In : Handbuch Policy-Forschung. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 1–28.

Jobin, Anna; Ienca, Marcello; Vayena, Effy (2019): The global landscape of AI ethics guidelines. In *Nat Mach Intell* 1 (9), pp. 389–399. DOI: 10.1038/s42256-019-0088-2.

Justo-Hanani, Ronit (2022): The politics of Artificial Intelligence regulation and governance reform in the European Union. In *Policy Sci* 55 (1), pp. 137–159. DOI: 10.1007/s11077-022-09452-8.

Kang, Cecilia (2023): Sam Altman, ChatGPT Creator and OpenAI CEO, Urges Senate for AI Regulation. In *The New York Times*, 5/16/2023. Available online at https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html, checked on 10/22/2023.

King, Gary; Keohane, Robert O.; Verba, Sidney (2021): Designing social inquiry. Scientific inference in qualitative research. New edition. Princeton, New Jersey: Princeton University Press.

Leslie, David (2019): Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.

Mahoney, James (2012): The Logic of Process Tracing Tests in the Social Sciences. In *Sociological Methods & Research* 41 (4), pp. 570–597. DOI: 10.1177/0049124112437709.

Manners, Ian (2002): Normative Power Europe: A Contradiction in Terms? In *JCMS: Journal of Common Market Studies* 40 (2), pp. 235–258. DOI: 10.1111/1468-5965.00353.

McCarthy, J.; Minsky, M. L.; Rochester, N.; Shannon, C. E. (1955): A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE. Available online at http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html, updated on 4/4/1996, checked on 10/22/2023.

Meseguer, Covadonga; Gilardi, Fabrizio (2009): What is new in the study of policy diffusion? In *Review of International Political Economy* 16 (3), pp. 527–543. DOI: 10.1080/09692290802409236.

Metz, Cade; Schmidt, Gregory (2023): Elon Musk and Others Call for Pause on A.I., Citing 'Risks to Society'. In *The New York Times*, 3/29/2023. Available online at https://www.nytimes.com/2023/03/29/technology/ai-artificial-intelligence-musk-risks.html, checked on 10/23/2023.

Mittelstadt, Brent (2019): Principles alone cannot guarantee ethical AI. In *Nat Mach Intell* 1 (11), pp. 501–507. DOI: 10.1038/s42256-019-0114-4.

Mittelstadt, Brent Daniel; Allo, Patrick; Taddeo, Mariarosaria; Wachter, Sandra; Floridi, Luciano (2016): The ethics of algorithms: Mapping the debate. In *Big Data & Society* 3 (2), 205395171667967. DOI: 10.1177/2053951716679679.

NIST (2019): A Plan for Federal Engagement in Developing AI Technical Standards and Related Tools in response to Executive Order (EO 13859) | NIST. With assistance of NIST. Edited by NIST. Available online at

https://www.nist.gov/artificial-intelligence/plan-federal-engagement-developing-ai-technical-standards-and-related-tools, updated on 10/21/2023, checked on 10/21/2023.

NIST (Ed.) (2021): Draft Taxonomy of Risks. With assistance of NIST. Available online at https://www.nist.gov/document/draft-taxonomy-ai-risk-october-15-2021, updated on 10/21/2023, checked on 10/21/2023.

NIST (Ed.) (2023): Artificial Intelligence Risk Management Framework. With assistance of NIST. Available online at https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf, updated on 10/21/2023, checked on 10/21/2023.

OECD (Ed.) (2023): Recommendation of the Council on Artificial Intelligence. With assistance of OECD. Available online at https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449#supportDocuments, updated on 6/13/2023, checked on 10/21/2023.

OMB (Ed.) (2020): Guidance for Regulation of Artificial Intelligence Applications. With assistance of OMB. Available online at https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf, updated on 10/21/2023, checked on 10/21/2023.

Papyshev, Gleb; Yarime, Masaru (2023): The state's role in governing artificial intelligence: development, control, and promotion through national strategies. In *Policy Design and Practice* 6 (1), pp. 79–102. DOI: 10.1080/25741292.2022.2162252.

Rességuier, Anaïs; Rodrigues, Rowena (2020): AI ethics should not remain toothless! A call to bring back the teeth of ethics. In *Big Data & Society* 7 (2), 205395172094254. DOI: 10.1177/2053951720942541.

Roberts, Huw; Cowls, Josh; Hine, Emmie; Morley, Jessica; Wang, Vincent; Taddeo, Mariarosaria; Floridi, Luciano (2023): Governing artificial intelligence in China and the European Union: Comparing aims and promoting ethical outcomes. In *The Information Society* 39 (2), pp. 79–97. DOI: 10.1080/01972243.2022.2124565.

Rogers, Everett M.; Singhal, Arvind; Quinlan, Margareth M. (2009): Diffusion of Innovations. In Don W. Stacks, Michael B. Salwen (Eds.): An integrated approach to communication theory and research. 2. ed. New York, NY: Routledge (Communication series Communication theory and methodology), pp. 432–448. Available online at https://www.taylorfrancis.com/chapters/edit/10.4324/9780203887011-36/diffusion-innovations-everett-rogers-arvind-singhal-margaret-quinlan.

Roose, Kevin (2022): The Brilliance and Weirdness of ChatGPT. In *The New York Times*, 12/5/2022. Available online at https://www.nytimes.com/2022/12/05/technology/chatgpt-ai-twitter.html, checked on 10/22/2023.

Schimmelfennig, Frank; Sedelmeier, Ulrich (2004): Governance by conditionality: EU rule transfer to the candidate countries of Central and Eastern Europe. In *Journal of European Public Policy* 11 (4), pp. 661–679. DOI: 10.1080/1350176042000248089.

Schmitt, Lewin (2022): Mapping global AI governance: a nascent regime in a fragmented landscape. In *AI Ethics* 2 (2), pp. 303–314. DOI: 10.1007/s43681-021-00083-y.

Sharp, Alexandra (2023): EU AI Act: European Parliament Passes Artificial Intelligence Regulation Bill. In *Foreign Policy*, 6/14/2023. Available online at https://foreignpolicy.com/2023/06/14/eu-ai-act-european-union-chatgpt-regulations-transparency-privacy/, checked on 10/21/2023.

Shipan, Charles R.; Volden, Craig (2008): The Mechanisms of Policy Diffusion. In *American Journal of Political Science* 52 (4), pp. 840–857. DOI: 10.1111/j.1540-5907.2008.00346.x.

Smuha, Nathalie A. (2021): From a 'race to AI' to a 'race to AI regulation': regulatory competition for artificial intelligence. In *Law, Innovation and Technology* 13 (1), pp. 57–84. DOI: 10.1080/17579961.2021.1898300.

Starke, Peter (2013): Qualitative Methods for the Study of Policy Diffusion: Challenges and Available Solutions. In *Policy Stud J* 41 (4), pp. 561–582. DOI: 10.1111/psj.12032.

The Economist (2023): The AI boom: lessons from history. Available online at https://www.economist.com/finance-and-economics/2023/02/02/the-ai-boom-lessons-from-history, updated on 10/23/2023, checked on 10/23/2023.

The White House (2021): US-EU Trade and Technology Council Inaugural Joint Statement. In *The White House*, 9/29/2021. Available online at https://www.whitehouse.gov/briefing-room/statements-releases/2021/09/29/u-s-eu-trade-and-technology-council-inaugural-joint-statement/, checked on 10/21/2023.

The White House (2022a): Blueprint for an AI Bill of Rights | OSTP | The White House. Available online at https://www.whitehouse.gov/ostp/ai-bill-of-rights/, updated on 3/16/2023, checked on 10/21/2023.

The White House (2022b): US-EU Joint Statement of the Trade and Technology Council. In *The White House*, 5/16/2022. Available online at https://www.whitehouse.gov/wp-content/uploads/2022/05/TTC-US-text-Final-May-14.pdf, checked on 10/21/2023.

The White House (2022c): US-EU Joint Statement of the Trade and Technology Council. In *The White House*, 12/5/2022. Available online at

https://www.whitehouse.gov/briefing-room/statements-releases/2022/12/05/u-s-eu-joint-statement-of-the-trade-and-technology-council/, checked on 10/21/2023.

The White House (2023): US-EU Joint Statement of the Trade and Technology Council. In *The White House*, 5/31/2023. Available online at https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/31/u-s-eu-joint-statement-of-the-trade-and-technology-council-2/, checked on 10/21/2023.

Tortoise (2023): The Global AI Index - Tortoise. Available online at https://www.tortoisemedia.com/intelligence/global-ai/#rankings, updated on 10/22/2023, checked on 10/22/2023.

Trampusch, Christine; Palier, Bruno (2016): Between X and Y: how process tracing contributes to opening the black box of causality. In *New Political Economy* 21 (5), pp. 437–454. DOI: 10.1080/13563467.2015.1134465.

TURING, A. M. (1950): I.—COMPUTING MACHINERY AND INTELLIGENCE. In *Mind* LIX (236), pp. 433–460. DOI: 10.1093/mind/LIX.236.433.

Ulnicane, Inga (2022): Chapter 14 Artificial intelligence in the European Union. In : The Routledge Handbook of European Integrations: Taylor & Francis. Available online at https://library.oapen.org/handle/20.500.12657/52622.

US Congress (2020): National Artificial Intelligence Initiative Act. With assistance of US Congress. Edited by US Congress. Available online at https://oecd.ai/en/wonk/documents/united-states-national-ai-initiative-act-of-2020-2020, updated on 10/21/2023, checked on 10/21/2023.

van Evera, Stephen (1997): Guide to methods for students of political science. Pbk. print., [Nachdr.]. Ithaca, NY: Cornell University Press (Cornell paperbacks).

Vogel, David (2009): Trading Up. Consumer and Environmental Regulation in a Global Economy: Harvard University Press.

Wendt, Alexander (1992): Anarchy is what states make of it: the social construction of power politics. In *Int Org* 46 (2), pp. 391–425. DOI: 10.1017/s0020818300027764.

Wendt, Alexander (1999): Social theory of international politics. 12. print. Cambridge: Cambridge Univ. Press (Cambridge studies in international relations, 67).

Weyland, Kurt (2009): The Diffusion of Revolution: '1848' in Europe and Latin America. In *Int Org* 63 (3), pp. 391–423. DOI: 10.1017/s0020818309090146.

Whittlestone, Jess; Nyrup, Rune; Alexandrova, Anna; Cave, Stephen (2019): The Role and Limits of Principles in AI Ethics. In Vincent Conitzer, Gillian Hadfield, Shannon Vallor (Eds.): Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. AIES '19: AAAI/ACM Conference on AI, Ethics, and Society. Honolulu HI USA, 27 01 2019 28 01 2019. New York, NY, USA: ACM, pp. 195–200.

Winfield, Alan F.; Michael, Katina; Pitt, Jeremy; Evers, Vanessa (2019): Machine Ethics: The Design and Governance of Ethical AI and Autonomous Systems [Scanning the Issue]. In *Proc. IEEE* 107 (3), pp. 509–517. DOI: 10.1109/JPROC.2019.2900622.

Winfield, Alan F. T.; Jirotka, Marina (2018): Ethical governance is essential to building trust in robotics and artificial intelligence systems. In *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* 376 (2133). DOI: 10.1098/rsta.2018.0085.