AI

MDPI

Article

From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems

Ellen Hohma * and Christoph Lütge

Insitute for Ethics in Artificial Intelligence, School of Social Sciences & Technology, Technical University of Munich, 80333 Munich, Germany; luetge@tum.de
* Correspondence: ellen.hohma@tum.de

Abstract: The current endeavor of moving AI ethics from theory to practice can frequently be observed in academia and industry and indicates a major achievement in the theoretical understanding of responsible AI. Its practical application, however, currently poses challenges, as mechanisms for translating the proposed principles into easily feasible actions are often considered unclear and not ready for practice. In particular, a lack of uniform, standardized approaches that are aligned with regulatory provisions is often highlighted by practitioners as a major drawback to the practical realization of AI governance. To address these challenges, we propose a stronger shift in focus from solely the trustworthiness of AI products to the perceived trustworthiness of the development process by introducing a concept for a trustworthy development process for AI systems. We derive this process from a semi-systematic literature analysis of common AI governance documents to identify the most prominent measures for operationalizing responsible AI and compare them to implications for AI providers from EU-centered regulatory frameworks. Assessing the resulting process along derived characteristics of trustworthy processes shows that, while clarity is often mentioned as a major drawback, and many AI providers tend to wait for finalized regulations before reacting, the summarized landscape of proposed AI governance mechanisms can already cover many of the binding and non-binding demands circulating similar activities to address fundamental risks. Furthermore, while many factors of procedural trustworthiness are already fulfilled, limitations are seen particularly due to the vagueness of currently proposed measures, calling for a detailing of measures based on use cases and the system's context.

Keywords: artificial intelligence governance framework; ethical duties; legal duties; AI ethics principle operationalization; responsible AI; semi-systematic review

Citation: Hohma, E.; Lütge, C. From Trustworthy Principles to a Trustworthy Development Process: The Need and Elements of Trusted Development of AI Systems. *AI* **2023**, *4*, 904–925. https://doi.org/10.3390/ai4040046

Academic Editors: Pablo Rivas and Gissella Bejarano

Received: 29 August 2023
Revised: 29 September 2023
Accepted: 10 October 2023
Published: 13 October 2023

1. Introduction

Numerous international governmental or non-governmental stakeholders have proposed fundamental principles for responsible AI that are supported by many organizations using and providing AI applications. A consensus on the fundamental values that shall build the foundation for responsible AI conceptualizations has been found around principles like transparency, justice and fairness, non-maleficence, responsibility, and privacy [1]. Ensuring that AI systems adhere to such fundamental system properties is expected to foster their perceived trustworthiness and increase stakeholder trust in AI technologies [2,3].

To incorporate these ethical principles in practice, a popular endeavor is to move responsible AI from principle to practice by operationalizing the derived characteristics for trustworthy AI applications. Mittelstadt [4], for example, confirms a lack of proven methods to translate AI ethics principles into practice, and Ryan and Stahl [5] argue that a mapping between higher-level principles and concrete methods is required to adopt them. Larsson [6] even more specifically concludes a "need for moving from principle

to process in the governance of AI" (p. 437). Particularly, duties that arise for the AI-providing organization are rarely concretely defined, although, of course, AI providers bear a central role in the AI actors' ecosystem and thus the move from principles for AI ethics to responsible AI in practice [7]. Based on the high-level fundamentals of trustworthy AI, AI governance research, therefore, has started to focus on elaborating implications of the defined principles to determine characteristics of responsible AI systems (e.g., [5,8]) as well as mechanisms to implement those (e.g., [9–11]).

This shows that the need for responsible AI is widely recognized, and the operationalization of abstract principles to concrete actions is often identified as an appropriate way to bring them into practice. However, the implementation of agreed-upon practices within the industry has not been as comprehensive as one might hope. This discrepancy can, among others, predominantly be attributed to two prominent obstacles frequently cited by practitioners. First, many organizations hesitate to take substantial action until comprehensive regulations are put in place [12]. Often, waiting for clear regulatory guidelines is preferred over additional efforts and burdens if one's own initiatives do not align with future provisions. Second, a lack of uniform, standardized approaches to translate the agreed guidelines to easily implementable and effective actions has been noted within the AI community [4,13,14], which is seen as a significant hurdle to their practical adoption. The diverse landscape of measures and guidelines proposed by various entities creates confusion and makes it challenging for organizations to discern the most appropriate path to follow. In addition, there is uncertainty about which specific measures are the most effective in ensuring trustworthy AI [12]. This ambiguity can inhibit decision-makers and make them reluctant to commit to any particular strategy.

Our article shall support practitioners in finding a solution to these problems. We approach the issue by moving the focus away from solely the trustworthiness of the product itself to a stronger focus on the perceived trustworthiness of the development process. While trustworthy AI is often defined as system requirements, in order to tangibly operationalize it, we need to understand its link to the measures along the development process. Therefore, we propose a concept for a trustworthy development process for AI systems. Our suggested framework is built off a semi-systematic literature analysis of AI governance efforts to derive obligations and measures to fulfill agreed AI ethics requirements and map them onto the AI development lifecycle. The results are compared to the implications for practical AI governance from the landscape of EU-centered regulatory frameworks.

Our research can support AI practitioners, particularly regarding the two major problems mentioned above. The review-based methodology shows a growing consensus regarding prominent measures of corporate AI governance. Incorporating well-known and diverse, action-oriented governance frameworks presents a unified summary and can provide clarity on which measures are generally proposed. The comparison with the EU-focused regulatory landscape shows a high degree of consistency between the proposed binding and non-binding measures and points to the core elements that can already be addressed without the final regulations. Finally, our proposed conceptual extension of trustworthy AI from solely a system configuration to the associated development process can be the first step towards a heuristic for determining the efficacy of a measure—the stronger it is linked to process trustworthiness characteristics, the stronger its potential for fostering stakeholders' trustworthiness perceptions, the underlying goal.

2. Theoretical Foundations of Trustworthy Processes for Operationalizing AI Governance

A fundamental goal for AI providers in operationalizing AI governance is to foster trust among their stakeholders. Societal actors are often predominantly impacted by the outcomes of AI, although they can only indirectly influence their system design. This results in a need for these stakeholders to trust in the responsible development of AI systems by the AI provider and other contributors and, thus, in return, their induced obligation to indicate their trustworthiness back to society.

To display the resulting necessity of establishing concrete trustworthy processes for AI development activities as well as requirements and first steps towards this, we outline the theoretical foundations of trustworthy processes in the context of AI governance research in the following. We start with an overview of the AI ecosystem, including its stakeholders and their power over AI development processes, to display the need for trust and, therefore, examine the underlying goal of trustworthy AI development. Building off traditional trustworthy software development concepts, we subsequently review opportunities for signaling trust in the development process and derive characteristics that can guide process trustworthiness assessment. Finally, we examine related research on the translation of AI governance mechanisms into trustworthy development processes to examine its current status and reasons for the problems regarding operationalization in practice.

2.1. The Need for Trust in the AI Ecosystem

A major authority when establishing trust lies with providers of AI systems, as they are responsible for understanding the user's requirements and translating them into technical applications. With their two key roles in realizing AI projects, deciding over and developing the system along with its main characteristics, the AI provider naturally holds a large share of power in the development process [7,15,16].

However, of course, it is intertwined and shared with a variety of different stakeholders that contribute to the development of AI systems at different stages. The different phases that are needed to evolve from the first problem statement to the final AI system deployment and post-processing are outlined in the AI development lifecycle. It typically includes (1) the problem understanding and design phase, where the problem, its characteristics, and requirements are determined and a solution drafted; (2) the data collection and handling phase, where relevant data are obtained, preprocessed, analyzed, and managed; (3) the model building phase, involving the actual model development and testing; and (4) the deployment and monitoring phase, where the system is deployed to the user and monitored over time [17–20]. Figure 1 summarizes the four stages of the AI lifecycle and their various required tasks and introduces the major stakeholders involved in each step.

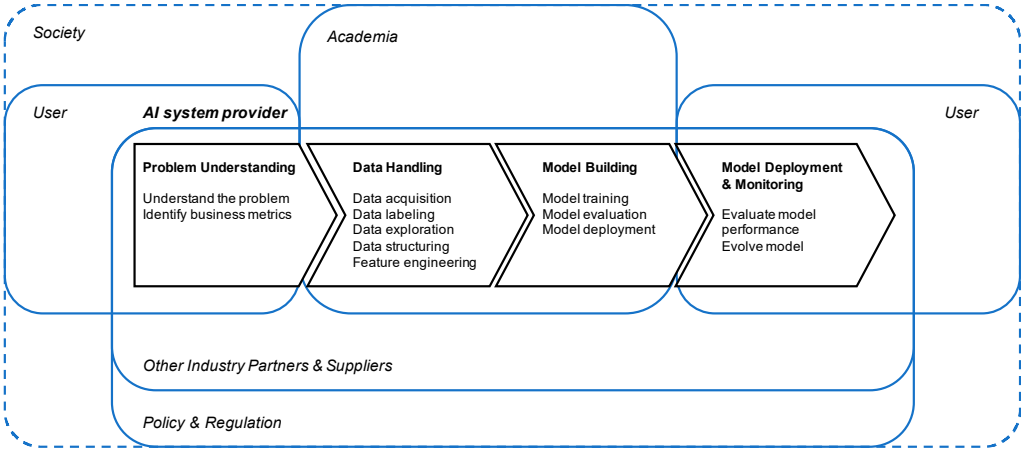

Figure 1. Stakeholder influence along the AI lifecycle.

The different stakeholder groups can influence different lifecycle stages to varying degrees. The user, representing the organization that operates the AI system after deployment, sets system requirements, supports problem understanding in the beginning, and engages and collaborates in deployment and after-monitoring at the end of an AI development project [10,21]. Apart from the system provider, this stakeholder group is also the one that can most actively control AI development. Other stakeholder groups, such as policy or academia, can consult, guide, or govern AI system development, e.g., through research, legislation, standards, and regulation; however, implementation of these guidelines is left to the AI system developers. Finally, the arguably most passively engaged

stakeholder group relates to broader society. They often require representatives, such as civil society communities that consult policy and industry, to enforce their demands in the AI development process [15].

Figure 1 shows that the AI ecosystem is manifold, and stakeholders can express and assert their interests with varying influence. In particular, the conflict that broader society is highly influenced by AI systems, however, can most passively engage in their development is still unresolved. This is confirmed by the many guidelines for responsible or trustworthy AI that place societal values at the center of considerations. Legislative efforts have further highlighted the high importance of fundamental and human as well as civil rights and democratic values (e.g., the AI Bill of Rights or the EU AI Act), prioritizing lawful, safe, and trustable AI applications [2]. However, societal engagement's outlined rather passive nature requires them to trust that these values are responsibly integrated into the design and development processes, which opens room for exploring how this trust can be promoted and fulfilled.

2.2. Characteristics of Trustworthy AI Development Processes

While concrete requirements and best practices for trustworthy development processes are continuously discussed in the more specific field of AI, their identification can draw from the common ground determined for general software development. With its main aim to reach trustworthy products, i.e., software that can satisfy objectives of trustworthiness based on predefined requirements [22,23], a trustworthy development process is the procedure through which such trustworthy products are created [23], i.e., the procedure by which the requirements for considering the outcome trustworthy are ensured. For software in general, characteristics of trustworthy products have been agreed upon and are often reported among security, privacy, reliability, or business integrity [24]. The development of trustworthy AI applications can draw on these characteristics; however, its enhanced capabilities require further adaptations. The consensus on the ethical fundamentals of trustworthy AI has led to the definition of requirements for trustworthy AI systems, often around the concepts of human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being, and accountability [2]. Most scholars from theory and practice argue that AI applications that fulfill these requirements can be deemed trustworthy.

This indicates that the foundations of system characteristics that trustworthy AI applications shall fulfill are already conceptualized. However, as pointed out earlier, ensuring trustworthiness only partially depends on the resulting system and, in addition, must take into consideration the process by which the system has been developed. Therefore, in order to evoke stakeholder trust, not only are the system characteristics that make an AI application trustworthy decisive, but also the characteristics that classify an AI development process as trustworthy. We can also benefit from the overlaps between AI development processes and regular software development. Although not finally settled, trustworthy development processes have been suggested and discussed for regular software. Standards like [25] on system life cycle processes for software engineering or [26] on a system security engineering approach have been assessed regarding their potential for introducing trustworthiness [27]. Further, standards like [28] or IEEE's approach to ethically aligned design [29] can give guidance on the central backbone of trustworthy development processes. The core goal in this endeavor is to enhance the predictability and controllability of development processes [23] and hence reduce the perceived dependability and uncertainty for the trustor. In particular, five characteristics are mentioned for development processes to be perceived as trustable. Figure 2 presents an overview of these characteristics.

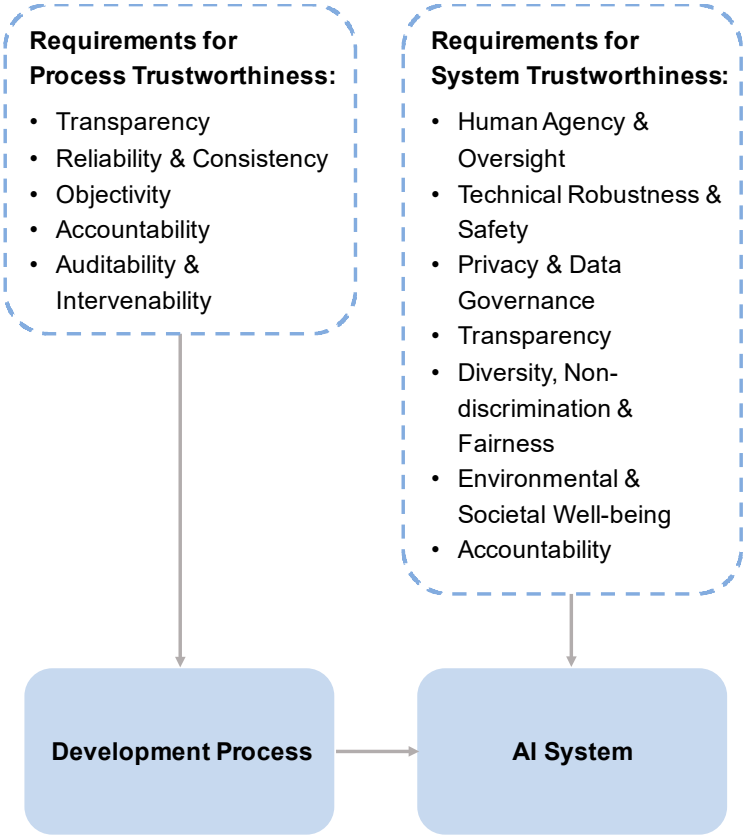

Figure 2. Requirements are presented that can be used to examine the trustworthiness of AI applications as outlined by the AI HLEG, as well as to examine the trustworthiness of the AI development process.

As a major pillar of trustworthiness, transparency requires clear and understandable communication about the steps taken during the AI development process, particularly making the process comprehensible and accessible to a wider audience, ensuring that people can understand how requirements for trustworthy systems are intended to be met. The steps taken within the development process to reach trustworthiness requirements must thus be clear, comprehensible, and communicable with a broader audience. The effects of transparency on trust have frequently been studied within and outside the information systems domain. Particularly within an AI context, transparency is often mentioned alongside explainability or interpretability, requiring that predictions or actions of AI systems are justifiable and traceable. The motivation behind transparency, however, differs according to the stakeholder and their perspectives. Weller [30], for example, lists eight types of goals that should be reached with transparency, ranging from transparency as a means for the developer to understand and debug a system to the facilitation of monitoring and auditing. Such explanations can foster the acceptance of the system and its design [31]. In a similar manner, transparency can facilitate acceptance regarding the planned development practices and is, therefore, a central factor for trustworthiness perceptions. On top, it lays the foundation for outside stakeholders to judge whether they perceive the envisioned steps as suitable for ensuring the responsible development of the system and thus have contributed to the prevention of undesired development practices.

Reliability and consistency are particularly important to enhance the process predictability, an essential component of fostering trust [32]. Hence, the development process must be reasonable, predictable, and standardized. A core goal is to prove to outside actors that processes follow a predefined plan and are not subject to arbitrary decisions or actions. At the same time, reliable and consistent development indicates that the goal of reaching responsible AI is pursued in a conscious and stringent way. A consistent process that is, at

best, transparently communicated can reduce uncertainty for the trustor. Obligatoriness can thereby further enhance process predictability. Mandatory activities in the process can ensure the user that certain minimum requirements have been fulfilled and clarify upcoming steps. Moreover, process credibility and legitimacy are supported if mandatory steps are based on a legal foundation.

Objectivity in this context refers to the absence of favoritism when processing data or deriving outcomes. It requires executing every step of the process in a neutral and standardized manner without adapting steps to certain preferred implications or outputs. All steps in the development process must be executed in an objective, unbiased, and fair way.

Further, concrete accountabilities determined for each step of the development process can foster user trust as they define responsibility and create the opportunity for more clearly tracing system malfunctions to obligations within the development stage. Clear obligations and accountabilities must, therefore, be discernible for the steps within the development process. Accountability is a core requirement for considering trustworthy AI or within legal frameworks. The EU AI Act mandates the definition of accountability frameworks for high-risk AI systems, and the HLEG defines it as one of the key requirements for trustworthy AI [2]. In a similar manner, it is a core characteristic of trustworthy processes. Paulus, et al. [22], for example, mention it as a means of fostering security in trustworthy software development.

Finally, the perceived trustworthiness of development processes can be enhanced through enabling external monitoring and control. Auditability and intervenability, i.e., monitoring, checks, and options for intervention from external stakeholders, are, therefore, crucial process requirements. The potential of using auditing practices to ensure ethical AI design has lately been widely discussed [33]. Thereby, auditing refers to the process of examining the consistency between a "set of auditable artifacts that record decisions, systems, and processes" [9] and stated principles, regulations, norms, standard metrics, or benchmarks [9,34]. Such checks are common for other high-risk technologies, such as in aerospace or finance, and have been found promising in the context of AI [35]. The fundamental definition, however, also shows that practices for auditing responsible AI cannot limit themselves to system properties but must consider the steps taken to ensure ethical design along the development. Therefore, options for outside checks and intervention are important characteristics for both the system and its development process when aiming to enhance trustworthiness.

2.3. Moving towards Trustworthy AI Development

Works from different actors guide activities in the move of responsible AI to practice and thus the development of corporate AI governance strategies, e.g., from policymakers, standards development organizations (SDOs), or research and academia. In particular, regulatory efforts are often demanded and seen as an appropriate way to unify currently proposed approaches [12]. Regulations such as the EU AI Act, with its proposed risk categories and mandatory risk-dependent countermeasures, can clarify important groundwork for AI providers and provide consistent guidance on what measures to take or which red flags to avoid. However, regulation can and should specify AI governance mechanisms only on a conceptual level. Going down to more specific, process-based provisions is clarified by accompanying industry standards. For example, the US National Institute of Standards and Technology AI Risk Management Framework defines important actions along the development of AI systems to govern, map, measure, and manage AI-related risks [36]. IEEE Standard 7000, on a standard model process for addressing ethical concerns during system design on a broader scale, outlines a process for system engineers to incorporate ethical values into their design practices [28]. The ISO/IEC JTC 1/SC 42 working group on artificial intelligence has (up to now) 20 published standards and 32 under development [37]. Among the published ones are fundamental groundworks on the trustworthiness of AI

(e.g., [38]) or AI risk management (e.g., [39]). More concrete guidance on, for example, how to integrate safety or transparency is currently under development.

Important work on operationalizing AI governance mechanisms also comes from research and the academic field. Here we see efforts within two, often interrelated streams: research proposing proactive approaches to set up responsible AI strategies or measures to actively implement those in the development process and reactive approaches referring to mechanisms to check—internally or externally—the responsibility of resulting systems or processes such as through auditing or impact assessment. Much research focuses on identifying and developing appropriate tools to integrate AI ethics into the design and development stage, particularly in proactive approaches. For example, overviews can be found on the current landscape of technical [8,10,11] and organizational [9,40] tools and methods. Important summaries on methods for responsible design, development, and deployment also come from the field of ethical AI assurance. Overviews and frameworks are proposed (e.g., [41,42]) and are slowly transitioning from mere theoretical considerations to actionable tools [43]. In contrast, reactive approaches focus less on the active adaptation of the development process and more on the evaluation of the resulting system. Nevertheless, they have important implications for the operationalization of responsible AI. Suggested impact assessments (e.g., [44,45]) can offer valuable guidance on the required system or process characteristics to implement. Auditing processes require the definition of "set[s] of auditable artifacts that record decisions, systems, and processes" ([9], p. 4) to allow checking them against predefined principles, metrics, or norms.

Therefore, while previous work already makes significant contributions to the operationalization of AI governance mechanisms, further efforts are needed to make them ready for practical application. Proposed tools can, at least from a theoretical perspective, solve many of the challenges; however, they are often considered unsuitable for their application in practice and are, therefore, rarely used [11,46,47]. In the field of AI assurance, Burr and Leslie ([41], p. 96), for example, specifically call for "practical systems and standards that can help teams and organizations" as a next research step. As a foundation for their adaptation to practical needs, a comprehensive understanding of the prominent measures is required to identify how they can support the development process.

With our article, we want to contribute to this transition to practice. The benefit and contribution of our research lie in integrating these approaches from the various stakeholder groups. Our primary objective is to provide a more comprehensive overview of the prominent measures and propel them into a tangible, actionable process to offer a unified, actionable summary—an asset that practitioners have pointed out as missing. This required a careful analysis of the multiple elements and their integration into a coherent framework. Finally, to measure its effectiveness, we examine the developed concept regarding its potential to facilitate the underlying goal: reflecting trustworthiness and, hence, fostering trust with stakeholders.

3. Methodology

Underlying our research is the assumption that in order to make an AI system trustworthy, we cannot only focus on the trustworthiness of the application itself but also take into account the trustworthiness of the development process. In this paper, our key objective is to derive the elements for a trustworthy process for responsible AI development from existing AI governance frameworks and discuss its potential for satisfying characteristics for process trustworthiness. We conceptualize the process from an analysis of regulatory and organizational frameworks that represent both legally binding and non-binding measures.

We have chosen a semi-systematic qualitative literature analysis methodology. This research method is used for heterogeneous topics that are conceptualized and studied by a wide variety of research disciplines [48]. In particular, it is needed where fully systematic reviews are impossible due to the complexity or variety of research approaches, topics, and types [49]. We follow this methodology, as we aimed for practice-oriented frameworks, including standards, civil society comments, or policy efforts, and thus, a systematic

keyword search on scientific search engines did not provide the anticipated results. Non-binding measures were collected from a variety of global AI governance frameworks and compared to legally binding measures that are currently enforced or planned in the EU.

3.1. Data Collection

Ethical and robustness-related obligations were obtained from four main fields of sources: (1) non-regulative policy efforts, (2) standards developing organizations (SDOs), (3) academic and research institutes or consortiums, and (4) non-governmental organizations (NGOs) or civil society groups. Actors to be investigated were retrieved from the aiethicist.org-repository [50]. The tabs "AI Principles" and "AI Governance" were systematically searched for stakeholders active in the field who had published documents giving insights on obligations and responsibilities for AI providers. As the primary focus of the study is how to operationalize trustworthy AI in practice, only documents that went beyond defining AI principles in general but included practice-orientated recommendations were included. A specific link to AI systems was required, i.e., documents where the primary focus lay on data governance were excluded. Further, these documents needed to be referenced on the respective stakeholder's website or found through regular online searches.

This systematic search resulted in a total of 155 considered stakeholders, of which documents on the responsible development of AI were found from 45 actors. Finally, 14 of these were found to have published practice-oriented reports on AI-related obligations, measures, or responsibilities, meeting the outlined inclusion criteria. Four documents were from non-regulative policy efforts [2,51–53], three were from SDOs [29,36,54], five documents were published by research institutes or academic consortiums [55–59], and two by NGOs or other civil society interest communities [60,61]. An overview of all retrieved stakeholders and those of which publications were included in the analysis can be found as Supplementary Material.

The main aim of the legal analysis was to provide an overview of thematic fields from which legal obligations for AI providers can arise when developing or deploying AI systems. The EU AI Act was used as a first point of reference to identify those. While it provides the backbone of AI-specific obligations, the EU AI Act's Explanatory Memorandum (particularly its sections 1.2 and 1.3) was used to determine related regulative fields. A context-independent search confirmed and extended a first draft of the categorization into obligation topics resulting from the AI Act's Recitals. For this, the EUR-lex summary repository [62] was used. Administered by the EU, it provides overviews of the main EU legal acts. The listed 32 policy fields were searched for EU decisions that could have an impact and result in obligations for AI providers. Finally, a non-systematic literature review on legal obligations for AI systems via Google Scholar confirmed the resulting five obligation fields presented in Section 4.1.

3.2. Data Analysis

The resulting documents were analyzed to retrieve the elements of the trustworthy AI development process. As shown in Figure 3, the identified policy and governance recommendations were used to derive obligations for AI providers from principles for trustworthy AI. These obligations were mapped to related measures to fulfill the identified duties. The resulting process for trustworthy AI development was compared to legally binding obligations from current EU law.

The non-binding obligations and related measures from policy and governance documents were retrieved using a thematic analysis methodology. The principles for trustworthy AI, more specifically, the seven key requirements for trustworthy AI as proposed by the AI-HLEG, were used as guidance for the analysis. For each principle, related obligations were determined iteratively by repeatedly working through all documents and retrieving codes. Similar to the process described by Braun and Clarke [63], these codes were then translated to more overarching themes, which can be found in the Supplementary Material.

The resulting themes were reviewed and refined until they represented distinctive AI provider obligations.

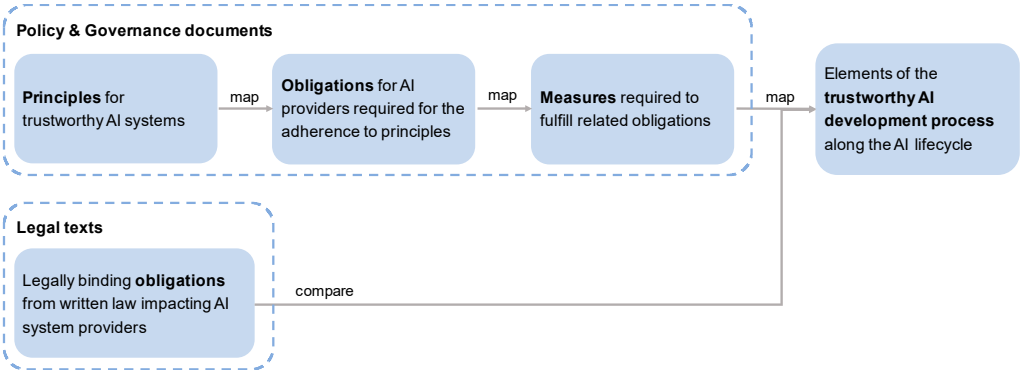

Figure 3. Procedure of mapping principles for trustworthy AI to a trustworthy AI development process.

In a similar manner, the documents were re-evaluated to retrieve the measures required to fulfill the identified obligations. For each identified obligation, related measures were derived through thematic analysis, retrieving and summarizing measure suggestions, and iteratively converging to a concise set of measures [64]. The resulting landscape of measures and the documents in which they were found can be reviewed in the Supplementary Material. Summarizing the resulting measures and mapping them along the AI lifecycle revealed the final elements of the trustworthy AI development process.

Finally, the resulting process was compared to obligations and measures suggested in binding legal texts. The identified regulatory fields and related EU-level legal documents were summarized regarding their imposed measures for the AI provider (see Supplementary Material). The resulting legally binding measures were compared to the developed trustworthy process, and elements of the process that were found legally enforced were marked accordingly. From this analysis, no new measures were added to the process, as no fundamentally new measures were found in the legal texts that directly impacted the providers' processes linked to the system's AI component.

4. From Trustworthy Principles to a Trustworthy Development Process

To investigate the elements as well as the state and implications of process trustworthiness in AI development procedures, we outline the derived framework for a trustworthy process of AI development from the existing fundamentals of trustworthy AI in the following. In particular, we elaborate on its foundation in the identification of AI provider obligations and the determination of related measures to address them.

4.1. From Principles to Obligations

The conducted analysis of obligations linked to responsible AI development supports the consensus on fundamental values that have been mentioned in the previous literature. Identified obligations are seen along the previously determined underlying principles. In Table 1, we report them along the seven key requirements for trustworthy AI due to their comprehensiveness and widespread acknowledgment. Within the columns, obligations are sorted according to the frequency with which they were found in the studied documents.

Table 1. Identified obligations for the AI provider along with AI HLEG's seven key requirements for trustworthy AI.

Human Agency and Oversight	Technical Robustness and Safety	Privacy and Data Governance	Transparency	Diversity, Non-Discrimination, and Fairness	Societal and Environmental Well-Being	Accountability
Ensure human autonomy/agency/determination Respect and protect fundamental/human rights Ensure human oversight Enable system termination Promote human augmentation	Ensure safety Ensure accuracy Ensure security Ensure reliability Ensure robustness Ensure validity Ensure reproducibility Ensure resilience to attack Ensure traceability Establish a fallback plan Ensure system quality Ensure verification	Ensure privacy Ensure data protection Ensure data quality Control data access Ensure lawful data processing Prevent data misuse/overuse Ensure data security Ensure data integrity Foster data risk awareness	Enable explainability of technical processes Communicate system capabilities and limitations Explain related human decisions/reasoning Ensure traceability of datasets and processes Inform about AI interaction Promote AI education Allow access for auditing Communicate intended use Ensure explicability Allow for intervention Ensure independence Ensure transparency on responsibilities Ensure truthfulness	Avoid/Correct/Monitor unfair bias Ensure non-discrimination Ensure diversity and inclusion Ensure equity, equality, and solidarity Ensure accessibility Ensure lawful development Enable multi-stakeholder engagement Enable compensation and remedy in case of discrimination Ensure peace and justice Define fairness Enable opportunity for correction	Prevent and reduce harm Monitor social impact Do more good than harm Ensure environmental friendliness Ensure proportionality to legitimate aim Ensure sustainability Monitor democratic impact Prevent misuse Establish multi-stakeholder dialog Ensure right foundation Ensure scientific foundation	Ensure auditability Provide documentation and information Assess general impacts Determine/assign responsibilities Allow for redress Establish appropriate oversight Establish ethics overseeing internal/external entity Establish measurement mechanisms Ensure public engagement Control access Foster accountability by design Create codes of conduct Collect feedback Ensure harm compensation

4.2. From Obligations to Measures

While the outlined obligations present the foundation for trustworthy AI development, moving closer towards implementation, measures to ensure their fulfillment can be derived. These measures were consolidated from existing AI governance frameworks and legal considerations with a focus on EU policy. The resulting list of binding and non-binding measures is presented in Table 2.

Table 2. Measures AI providers can adopt to fulfill their obligations according to the studied AI governance documents.

	Non-Binding	Binding *
Plan	Create **codes of conduct**	Develop AI **governance strategies** regarding: - trustworthy AI measurement and evaluation - data protection and access - quality management - risk management - human intervention - displacement and business change - privacy and accountability (by design) - education and awareness raising regarding harms and system misuse Determine/assign **responsibilities and accountabilities.** Set **requirements and thresholds** for: - system safety - accuracy, reliability - quality of data preparation and training - supporting hardware, software (incl. cloud applications) - industrial and consumer use case **Redress and compensation** for harms (incl. due to discrimination)

Table 2. *Cont.*

	Non-Binding	Binding *
Create and establish	Establish **participatory development processes** through: - pull mechanisms: offer public feedback opportunities, adoption of open standards, and interoperability to facilitate collaboration - push mechanisms: clarification of public concerns and questions, consultation of directly or indirectly affected stakeholders Create **ethics overseeing** internal/external entity Ensure **team diversity** regarding backgrounds, cultures, disciplines Establish **risk prevention/management** regarding: - wrong, unintended, or forbidden use of data - data modification or abuse - fairness-related harm - adversarial patch tricking - human errors - intentionally or unintentionally included biases Avoid, correct, and monitor **unfair bias** through: - removing identifiable discriminatory bias where possible - testing and monitoring mechanisms - evaluating how potential biases might arise Enable and ensure **human control** over data and processes **Educate** relevant personnel Ensure **oversight and control** regarding: - system purpose, constraints, requirements, decisions - shut down or modify misbehaving systems Enable **auditing** through: - developing audit trail requirements - provide access for internal or external auditing	Apply **systematic risk management** (incl. a fallback plan) Enable **human oversight** (human-on-the-loop) or **human control** (human-in-the-loop) by the user to: - assess and rectify incorrect predictions - avoid human subordination - avoid basing decisions with significant impact solely on automated processing - enable attribution of ethical and legal responsibility - terminate the system if human control of the system is no longer possible Provide options for **public intervention and participation** regarding: - choosing which digital services to use or to avoid using them - correcting false information - questioning and changing unfair, biased, or discriminatory systems - right to a final determination made by a person - consider bias and safety bug-bounty programs Ensure **explainability of technical processes**, e.g., through using tools, regarding: - system outcomes (incl. why similar-looking circumstances generate different outcomes) - logic or algorithm behind the outcome - main factors in a decision - data quality, accuracy
Assess and evaluate	Ex ante **impact assessments** regarding: - fundamental and human rights - privacy - society and societal norms - sustainability and environment - democracy - system's legitimate, proportionate, and scientific foundation Evaluate opportunities for quality labels and **certifications** Evaluate **independence of (critical) infrastructure** Ex post **impact assessment** regarding: - system accessibility - unfair denial of resources, rights, goods, participation	Assess compliance with applicable international and domestic **legislation, standards, and practices** **Test data quality** regarding: - accuracy - actuality - integrity - representativeness Assess and ensure **lawfulness** of data processing regarding: - protection of data and metadata - data access and control - user's freedom of intrusion - limiting observations **Test** system regarding: - accuracy/reliability (through model selection, measurement metrics, mitigation of model over-/underfitting) - robustness (through sensitivity analysis) - security (regarding data poisoning, model leakage, unexpected, adversarial or malicious use, cybersecurity threats) - discrimination (regarding use of protected classes and dataset representativeness) - validity - verification - domain-specific requirements (through simulation, in-domain testing, software/hardware requirements) Assess and ensure **lawful** development Assess and ensure **truthfulness** regarding statements to customers and consumers
Document and communicate	Support **AI education** through: - supporting educational curricula and public awareness activities - engaging with civil society to understand best practice for education	**Documentation** and record-keeping of: - datasets - data provenance, gathering, and labeling - data testing processes - data access **Document** for traceability and reproducibility: - provide replication files - use tools (e.g., to abstract computational graphs and archive data at each step of transformation pipelines) - adopt open standards **Disclose during use**: - interaction with AI system - which consumer actions can negatively impact scores/decisions **Communicate** with relevant stakeholders, e.g., in the form of use manuals: - definitions of key concepts and measures - system purpose, reach, (intended) use, capabilities, and limitations - design decisions, including characteristics of training data and model structure - data protection processes - responsible internal and external actors Provide **documentation** of: - system goals - design choices including definitions, standards, testing, measurements, and assessments of performance, privacy, and fairness - identified biases and their potential impacts - stakeholders involved and their responsibilities - risk management structures, including monitoring, feedback, error correction, human control, and cybersecurity

* Binding measures were found in the studied EU regulations or directives as mandatory for some AI technologies or under certain conditions.

4.2.1. Measures According to AI Governance Documents

Measures to address the obligations of AI providers were found, as similarly suggested in previous work by Mäntymäki, et al. [65], among activities for planning, assessment, and ensuring creation and communication.

Planning activities include mechanisms for establishing the strategic alignment of system development and associated risks regarding the AI provider's obligations, including the creation of governance strategies and codes of conduct, frameworks for risk management and accountability attribution, as well as determination and documentation of system requirements and thresholds. A fundamental goal to foster is the prevention of harms related to violation of rights, privacy, safety, security, sustainability, or, generally, the public good [52]. Considerations for strategy development should be based on conditions

of normal use and potentially unanticipated but foreseeable use or misuse [51]. In addition, the diversity of cultural norms within the user groups should be taken into account, whereby the inclusion of different stakeholders during the creation of strategic directions can help [29].

Mechanisms to assess the fulfillment of certain AI provider obligations and ensure appropriate action-taking if needed include standardized impact assessments (including technical testing) before and after development on certain system properties and implications, as well as evaluation of system dependencies, lawfulness of data processing, and development, team characteristics and capabilities, options for auditing (including quality labels and certifications), truthfulness, and ensuring of compensation. The rationale behind this group of measures is to ensure that AI systems should only be deployed after a proper assessment of their purpose, objectives, benefits, and risks [61], bearing in mind that these assessments must be proportionate, rational, and methodologically sound [2].

The group of mechanisms to establish certain conditions comprises activities linked to obligations that require or impact the active (re-)design of the system or organizational processes linked to it. This is mainly required regarding the creation of participatory development and public intervention mechanisms, measures to ensure human oversights and control (including a focus on monitoring ethical aspects), as well as provision of certain documentation and enabling of explanation to ensure that processes can be appropriately steered. It is important for the implementation of such measures to take into account "shifts in technologies, the emergence of new groups of stakeholders, and to allow for meaningful participation by marginalized groups, communities and individuals" [52].

Finally, communication activities are a prerequisite for multi-stakeholder engagement. Such measures are linked to the disclosure of certain information, communication of system definitions, purpose, limitations, risks, and use and education of staff, users, and the general public. Particular proactive engagement from AI providers has been demanded regarding the evaluation of augmenting human capabilities, advancing inclusion of underrepresented populations, reducing economic, social, gender, and other inequalities, and protecting natural environments [51]. For due and effective communication measures, the information provided needs to be understood and accurate and disclosed to the general public or the responsible human in charge [60].

4.2.2. Comparison to Legal Perspective

Requirements from legal frameworks have been found among five thematic fields and depend on the type and use of an AI-based application. Figure 4 presents an overview of the identified and analyzed regulatory texts.

To account for risks that arise specifically from the development and use of AI, the Regulation of the European Parliament and the Council for Laying Down Harmonized Rules on Artificial Intelligence and Amending Certain Union Legislative Acts, or short, the AI Act [66], suggests a four-level risk categorization into unacceptable, high, limited, and minimal risk systems. AI systems bearing unacceptable risks will be prohibited partially or in their entirety from use on the EU market; high-risk systems will be subject to conformity assessments, and thus, further restrictive measures will be implemented, mostly for the AI provider. While systems posing limited risks will need to ensure certain transparency requirements, minimal-risk systems are free of binding measures, only recommended to provide and follow self-imposed codes of conduct to voluntarily commit to the same response measures as high-risk systems.

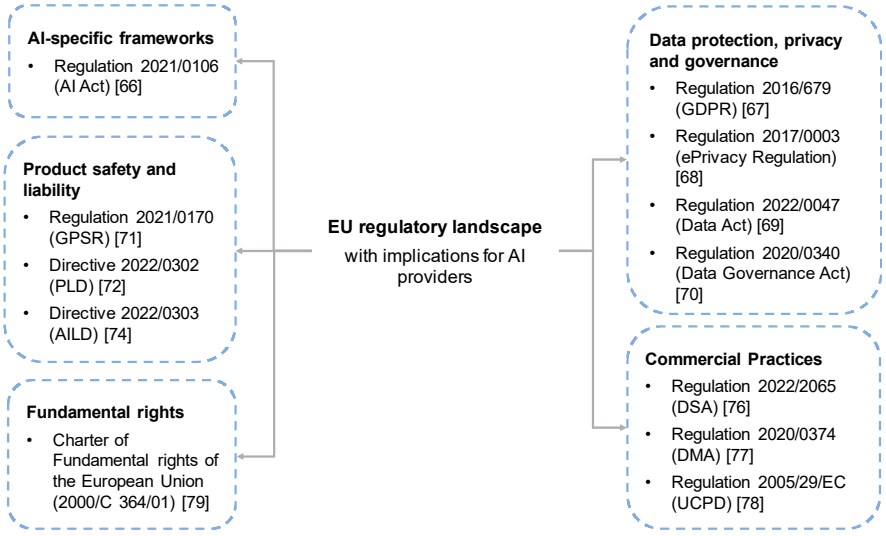

Figure 4. Overview of the identified primary fields of the EU regulatory landscape that have impacted the responsibilities of AI providers.

As AI systems rely on data processing, they are subject to the restrictions set out by the General Data Protection Regulation (GDPR) [67], particularly when it comes to processing personal data or even special categories thereof. In such cases, the processing must meet additional safety and privacy requirements. In addition, special rights are granted to the data subject, which allow them to demand certain handling of their data from the data operator. While the GDPR, thus, directly entails mandatory measures for the AI provider, there are further EU regulations concerning data, both personal and non-personal, that affect the AI provider, however, only to a limited extent. Regulation 2017/0003 [68] on Privacy and Electronic Communications (ePrivacy Regulation), provides more concrete specifications for electronic communications data and thus complements GDPR, where such data qualify as personal. It generally restricts interference with electronic communications data, e.g., in the form of listening, tapping, storing, or monitoring, to certain permitted use cases (ePrivacy Regulation, Art. 5). The proposal for Regulation 2022/0047 [69] on harmonized rules on fair access to and use of data (Data Act) sets out obligations for the provision of data generated by the use of (physical) products that collect and transmit data. Regulation 2020/0340 [70] on European data governance (Data Governance Act) regulates the reuse and sharing of data between stakeholders in the EU to strengthen data availability and exchange. While these legislations may impact the AI-providing organization, e.g., if the system falls within one of the covered use cases or if the AI provider is involved in data collection and post-processing, their set out obligations, however, are linked to the overall system and have less direct implications for the AI component.

If AI components are treated as or built into products, requirements from product safety and liability can impose obligations and, therefore, require protective measures from the AI component provider. In this regard, the two constructs of the General Product Safety Directive (GPSD) [71] and Product Liability Directive (PLD) [72] provide two complementary mechanisms for enforcement of damage-related consumer claims, where the PLD outlines the liability specifications to assert claims that result from a defect or unsafe product, and product safety regulations lay down the specifications that a product must adhere to in order to be considered safe. While the AI Act outlines some safety-related AI requirements, the new GPSR "provides a safety net for products and risks to health and safety of consumers that do not enter into the scope of application of the AI proposal" (GPSR), and therefore AI-equipped products that are not subject to the more specific safety rules of the AI Act, i.e., products where the AI component is considered to pose only minimal risk, must comply with the provisions of the GPSR [73]. Nevertheless, the legislator sees a particular need for action regarding AI and supports the PLD with the proposition of

an AI Liability Directive (AILD) [74] that explicitly addresses claims in relation to AI-based systems. Here, especially access to information on high-risk systems and the burden of prove shall be adapted to the specific circumstances of AI (AILD, Art. 1(1)).

In order to "create a safer digital space where the fundamental rights of users are protected and to establish a level playing field for businesses" [75], the EU has adapted and extended some of its existing regulations in the area of commercial practices to further strengthen consumer rights and trust. This particularly includes the development of Regulation 2022/2065 [76], the Digital Services Act (DSA), and Regulation 2020/0374 [77], the Digital Markets Act (DMA). The DSA sets restrictions for digital services providers, defined as providers of "intermediary service", such as conduit, caching, or hosting services. AI-based use cases are affected in many ways, predominantly in the form of online advertisement targeting and recommender systems. The DMA similarly imposes additional obligations for "gatekeepers", i.e., large providers of core platform services, such as Amazon, Apple, Google, Meta, or Microsoft. While most obligations relate to restrictions of limiting access to data and services, AI-based applications are particularly addressed when it comes to recommender systems. In addition to the specific rules of the digital domain, existing legal frameworks on commercial practices also carry responsibilities for the AI provider. Directive 2005/29/EC [78], the Unfair Commercial Practices Directive (UCPD), prohibits misleading and aggressive commercial practices. Particularly, restriction of misleading practices can impact AI applications, as the AI provider must not provide false information or deceive consumers regarding, among others, the main characteristics of products, such as their benefits and risks, and compliance with promoted codes of conduct.

Finally, human or fundamental rights, founded in the Charter of Fundamental rights of the European Union (2000/C 364/01) [79], are the foundational basis for most of the above-outlined legal frameworks. In the implementation of Union law, they apply between EU institutions and bodies and the people; therefore, in the context of AI, no direct obligations for AI providers can be inferred from the legal text. Nevertheless, they are frequently mentioned in the context of AI, as risks and challenges for the respect of human rights are often identified with the introduction of AI systems [80]. This results in a direct obligation for the state to protect its citizens from restrictions on fundamental rights and an indirect obligation for the AI provider to comply with the stipulated provisions and measures against imposing restrictions on fundamental rights.

In summary, the policy takes a similar view to AI governance and has already incorporated some of the determined measures into regulatory and legal guidance. The conditions under which they are mandatory depend on the particular AI application, e.g., a systematic risk management procedure is required only for AI systems classified as high-risk under the EU AI Act, or the engaged target group, e.g., truthful statements regarding a system's properties and capabilities are particularly mandatory for communication with consumers according to UCPD. Therefore, we would like to note that our indications in Table 2 regarding binding practices should not be read as legal advice on which mechanisms to implement but constitute an analysis of which methods are regarded as both relevant and generalizable enough by legislators to be mandated across multiple use cases. This objective can provide interesting insights when examining which measures are suggested as mandatory or not and is useful for the development of a standardized, trustworthy AI development process.

4.3. From Measures to Process

The determined measures result in a framework for the trustworthy development of AI systems. Figure 5 outlines the derived activities and outputs.

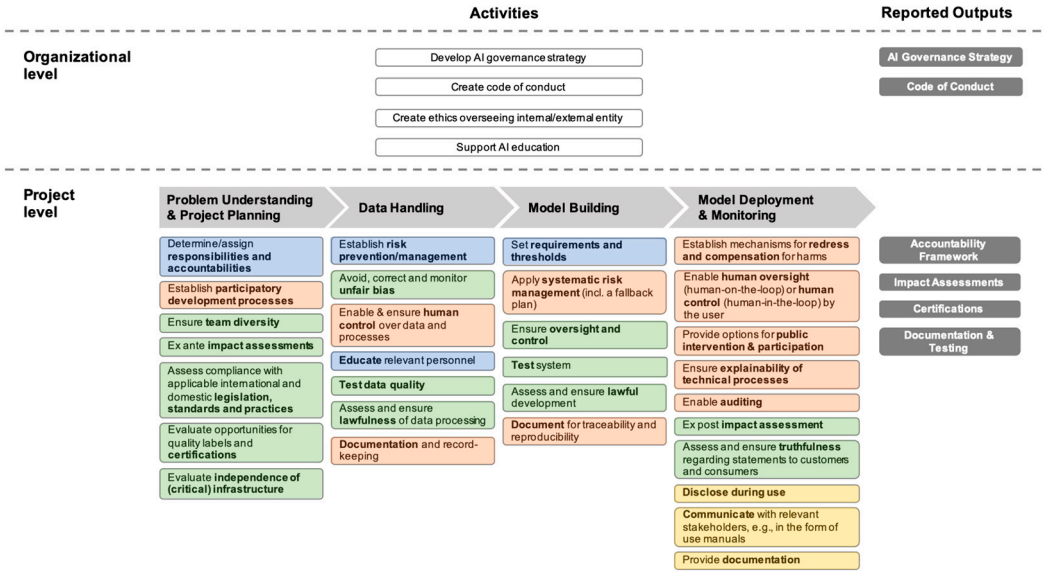

Figure 5. Core framework of a trustworthy development process for AI applications. Colors indicate measures to plan (blue), assess and ensure (green), create (orange), and communicate (yellow).

Measures within the process were found on two levels: activities that relate to strategic decision-making and, therefore, must be clarified on an organizational level and activities that depend on the use case or context-specific circumstances and, therefore, are specified per project. Organizational-level activities include strategy-setting tasks, such as developing a corporate AI governance roadmap or planning AI education and training offerings within and outside the organization. Project-level activities are structured along the AI development lifecycle, outlining the tasks that are required in each stage according to our analysis. Reporting certain outputs within or outside the organization is recommended on both levels.

5. Discussion

The proposed conceptual framework of a trustworthy development process for AI applications outlines the measures that have been identified to ensure the trustworthiness of AI applications regarding their responsible development and use. The identification of prominent AI governance measures from regulatory, policy, standardization, and research activities can support AI providers in setting up appropriate corporate AI governance strategies. In the following, we discuss its potential for fostering outlined process trustworthiness characteristics as well as determine the next steps in the practical implementation of such activities and processes.

5.1. Process Trustworthiness of Current Responsible AI Development

Previous research has indicated that trustworthiness can be enhanced by facilitating stakeholders' perception of certain key process characteristics. Transparency, reliability and consistency, objectivity, accountability, and audibility and intervenability were seen among the most crucial ones for procedural trustworthiness. To foster those, currently proposed measures for trustworthy AI and, more specifically, the derived process for trustworthy development of AI systems can show promising potential, above all by two important properties of the derived development process.

First, the clear structure it introduces can offer stakeholders insight into the measures that will be taken to prevent ethical problems or irresponsible AI. The implementation of a clear and structured framework to address responsible and trustworthy AI has become essential for stakeholders in the field; however, clarity regarding required measures is often mentioned as a major drawback [12]. The derived development process outlines a series of concrete steps and actions that can be taken at different levels to prevent and

mitigate potential ethical challenges and the proposed outcomes to be communicated more broadly. This can have an impact on transparency on several levels, firstly through the fact that there is a concrete process, which in the best case is also openly communicated, and secondly, more obviously, through the proposed publicly reported outputs. In that, it promotes two important functions of transparency. The clear structure of measures helps stakeholders understand a system's strengths and limitations, and the clear communication of processes and outputs allows for checking whether a system works appropriately and, in this case, ethically [30]. In addition, its second function, the disclosure of processes and thus comprehensibility of mitigating measures, influences the perception of auditability. As previously outlined, such auditing processes require artifacts and evidence to be checked against the established ethical principles [9,34]. The determined development measures make it easier to meet this requirement, as the artifacts needed for auditing can be derived from the concrete actions taken throughout the process. Trustworthiness can further be enhanced through the introduction of legitimacy. The proposed measures result from the analysis of existing framework conceptions; most have been subject to a consultation period involving relevant stakeholders in the framework's development. This ensures that the decisions made are well-founded and align with the values and expectations of the broader community. Finally, obligatoriness is signaled in the process. The derived development process identifies certain steps as mandatory, reinforcing its commitment to ethical practices and responsible AI. This emphasizes the seriousness of the endeavor and underscores the importance of adhering to these guidelines in AI-related initiatives.

A second key property of the derived development measures that can influence perceived procedural trustworthiness is the breakdown of steps into specific tangible actions. Such a step-by-step approach can foster accountability by breaking an AI provider's ethical obligations down to related countermeasures, enabling the definition of responsibilities on this level. This, in return, further enhances an essential function of explanation and, thus, a key requirement for trustworthy AI, enabling the meaningful challenge of an AI system's outputs [30].

However, while such an approach can thus ensure much improvement of procedural trustworthiness, there are also certain characteristics in the current form of the process that demand further reiteration to fully support trustworthiness perceptions.

One significant issue that pertains is the vagueness of some steps within the process. While many steps are well-defined and offer a clear roadmap for addressing ethical concerns, some remain ambiguous. This lack of preciseness can hinder perceptions of the reliability of the overall process. If certain steps are open to interpretation, it might be left to individual organizations or decision-makers to decide how they are implemented. This subjectivity could lead to inconsistent practices and potentially compromise the overall effectiveness of the process. Moreover, the binding nature of the procedure is not always clear. Even in the investigated legally mandated measures, there is some leeway on the interpretation and requirements for implementation. The level of obligatoriness might vary depending on the specific use case, context, or area of implementation. Such inconsistencies could raise questions about the enforceability and effectiveness of the process. Stakeholders might wonder whether the process's guidelines are universally binding or if they can be overlooked or circumvented in certain circumstances, leading to potential compromises in the perceived trustworthiness of the process.

A final point of consideration when evaluating the process's effect on trustworthiness is its ability to actually ensure the system characteristics that are seen as required for trustworthy AI. The question can be raised whether merely following the prescribed steps is sufficient to ensure a trustworthy AI system. For example, while we have seen that the clarified steps can promote some functions of transparency, others, such as the need to "overcome the reasonable fear of the unknown" [30], are not necessarily addressed. The "unknown" is often related to malfunctioning or misuse of the system, but it is not necessarily clear whether the proposed measures are sufficient to avoid this in the best possible way. The consensus identified on certain measures to ensure the trustworthiness

of AI systems and the high level of attention currently being paid to this issue may indicate that at least all the obvious measures have been defined. However, whether these measures can finally lead to enhanced trustworthiness of the systems and more stakeholder trust in the development procedure is neither clear for procedural trustworthiness nor is it solved in general.

5.2. Next Steps in Trustworthy AI Development

Our analysis indicates a high-level consensus around measures that can be taken to mitigate ethical issues of AI systems along the development process. Clarity is often mentioned by practitioners as a major drawback to the realization of AI ethics principles. While the entirety of developed concepts can bring clarification on what is generally required from AI providers to ensure responsible AI, the vagueness of some proposed measures is also apparent in our analysis. Particularly when breaking obligations down into respective measures, we see that while some measures seem to be quite "ready-to-use", others require further efforts to apply them to real-world scenarios. For example, while the task of determining and assigning responsibilities seems straightforward, recent studies have indicated that accountabilities for AI systems are often ambiguous in reality, calling for the creation of detailed accountability frameworks [12,81,82]. In implementing such development processes, it is important to identify with practice which of the measures are already implementable and which need more detailing. The context will certainly play a major role here, as different industry use cases have different characteristics and requirements regarding the demand for and magnitude of obligations [83]. We see that legislation already accounts for this, and, for example, the current proposal of the AI Act classifies AI systems according to their use case industry into risk levels. The scope of imposed, binding measures followingly depends on the risk that results from a system's use. While our analysis gives an overview of which measures exist in general, we make no assumptions about which of them are relevant or more important than others given certain contexts. Clearly, a simultaneous implementation of all the measures described does not seem realistic or desirable for all AI systems. Future research can, therefore, focus on how the summarized obligations can be applied to real case studies and what this implies for the trustworthiness of AI systems.

Further, while recent research often focuses on the technical implementation of responsible AI, for example, through tools or "by-design" concepts, our analysis shows that AI governance mechanisms currently are largely recommended among non-technical methods, for example, regarding strategy-making, documentation, and communication. This either reemphasizes the inability of available technical tools to meet the needs of practice or suggests that careful consideration must be given when automated or "by-design" approaches are feasible and desirable and when a non-technical assessment of activities, perhaps based on human intuition, is required. For instance, although technical tools to detect bias in training data might be helpful, a thorough interpretation of the results regarding whether the identified biases result from unfair assumptions will surely be needed. The notion that AI governance measures are largely non-technical methods further impacts the current view on accountability for AI systems. In comparison, system developers and designers are often named as responsible entities for ensuring responsible AI [81]. Most of the identified measures are not directly linked to system implementation and would require inputs or action from further departments, such as those related to strategy, communication, and management. This suggests that a more concrete examination of what actions should be taken to fulfill which obligations, taking into account the context of application and the characteristics of the expected mechanisms, can further facilitate the identification of the bodies or roles that can be held accountable for the outcomes of an AI system.

Finally, our analysis presents a comprehensive list of obligations and measures that current, practice-oriented AI governance frameworks offer to ensure the trustworthy behavior of AI. It, however, cannot answer the question of whether implementing these measures, to a reasonable extent and with reasonable effort, will finally increase stakeholder trust and

lead to a proper realization of the goal of responsible AI. Trust is a property of an individual trustor, who, based on personal perceptions and experiences, beliefs that an outcome is beneficial enough to engage in an unknown situation [22,84]. Being a psychological state of the trustor based on their subjective perceptions and decisions [85], trust can only be influenced; however, it cannot be directly controlled by the trusted organization. Both theoretical and empirical research suggest that certain measures to implement ethical considerations into system design and hence to signal trustworthiness positively influence the stakeholder perceptions—for example, measures to enhance fairness and individuality can promote user satisfaction with applications [86] or certain system design choices such as human-in-the-loop architectures can help reduce algorithmic aversion [87]. Nevertheless, the final decision remains to the respective stakeholders. Therefore, in practice, appropriate techniques for measuring the impact achieved, as well as continuous monitoring and re-evaluation of the measures implemented, will be key.

5.3. Limitations

Our derived framework of a trustworthy AI development process provides a unified overview of corporate AI governance mechanisms as proposed by various stakeholder groups, and the resulting clarity can thus support AI providers in the development of responsible AI governance strategies. However, there are also limitations to the scope of our results.

The chosen methodology was found suitable and required for determining practice-oriented recommendations from a variety of stakeholders. However, due to the semi-systematic approach, it is possible that other similarly relevant documents were not considered. In addition, given the rapid growth of this field, further practical guidelines may emerge which have not been included in this research. However, given the comprehensive approach to the identification of stakeholders, the diversity of considered stakeholder groups, and the similarity of identified measures as determined in the analysis, we do not see this as a weakness of our study. A second limitation stems from the required geographical focus when analyzing the implications of regulation. It was necessary to limit the scope either in breadth or in depth, which is why we opted for a granular review only of the EU regulatory landscape. However, given the EU's leadership in responsible AI governance and their advanced regulation in this field, we regard this as a minor limitation to our analysis and, in contrast, see valuable insights for other geographical areas.

6. Conclusions

The need to detail principles of ethical AI and adapt them for operationalization in practice is repeatedly emphasized in various fields. Roads to this goal are seen in assessing the current governance landscape, further clarification, and detailed conceptualizations [4,13]. Particularly from a practitioner's perspective, unification and clarification regarding responsibilities and related measures are needed to support them in establishing appropriate corporate AI governance strategies.

Our research aims to support these required objectives by advancing research on trustworthy development processes for AI applications. We explored the essential characteristics that define a process as trustworthy, drawing upon traditional concepts from trustworthy software development. By investigating the fulfillment of these trustworthy development propositions within frequently proposed trustworthy AI measures, we assessed whether they can effectively ensure responsible AI applications. Through a semi-systematic literature analysis of AI governance efforts and EU-centered regulatory frameworks, we translated agreed AI ethics requirements into practical obligations and derived the measures suggested to fulfill them. By mapping these measures onto the AI development lifecycle, we conceptualized the framework of a trustworthy AI development process.

Our research can, therefore, provide important insights for the practical implementation of AI governance measures. Obligations of AI providers to comply with the agreed ethical principles of AI were determined, and the corresponding measures that can be

implemented to fulfill these were systematically retrieved. The resulting concept of a process for the trustworthy development of AI systems can help support clarity on the state of the art of demanded mechanisms. Finally, the discussion on the degree of process trustworthiness that can be fulfilled with such a process sheds light on the overall state of trust in AI applications.

While our analysis can thus provide much clarification regarding the steps toward principle operationalization, it also sheds light on the open questions that will be clarified next. While a general catalog of measures applicable across various application scenarios seems useful to obtain a standardized overview, a case-based consideration is needed to identify which obligations and related measures are seen as particularly relevant for certain uses and, more generally, how to determine a heuristic to discover these variances. Further, a clearer distinction in which use cases or contexts automated technical approaches are feasible and desirable, and thus, developing technical tools to implement them is needed. Finally, our results provide a comprehensive overview of the tasks required to fulfill certain AI provider obligations. Whether these tasks are reasonable in practice and particularly whether they are enough to thoroughly consider an AI system as appropriately responsible might require further measurement mechanisms or independent assessments.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ai4040046/s1, Table S1: Data Sources; Table S2: Obligations and Measures retrieved from the Governance Documents Analysis; Table S3: Requirements and Obligations retrieved from the Legal Analysis.

Author Contributions: Conceptualization, E.H.; Methodology, E.H.; Formal analysis, E.H.; Investigation, E.H.; Resources, E.H.; Writing—original draft, E.H.; Writing—review & editing, E.H.; Visualization, E.H.; Supervision, C.L.; Project administration, E.H. and C.L.; Funding acquisition, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Fujitsu Limited and the Technical University of Munich's Institute for Ethics in Artificial Intelligence (IEAI).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
2. High-Level Expert Group on Artificial Intelligence (AI HLEG). *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019.
3. Bartneck, C.; Lütge, C.; Wagner, A.; Welsh, S. *An Introduction to Ethics in Robotics and AI*; Springer Nature: Berlin, Germany, 2021.
4. Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nat. Mach. Intell.* **2019**, *1*, 501–507. [CrossRef]
5. Ryan, M.; Stahl, B.C. Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. *J. Inf. Commun. Ethics Soc.* **2020**, *19*, 61–86. [CrossRef]
6. Larsson, S. On the governance of artificial intelligence through ethics guidelines. *Asian J. Law Soc.* **2020**, *7*, 437–451. [CrossRef]
7. Deshpande, A.; Sharp, H. Responsible AI Systems: Who are the Stakeholders? In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES 22), New York, NY, USA, 7–8 February 2022; pp. 227–236.
8. Georgieva, I.; Lazo, C.; Timan, T.; van Veenstra, A.F. From AI ethics principles to data science practice: A reflection and a gap analysis based on recent frameworks and practical experience. *AI Ethics* **2022**, *2*, 697–711. [CrossRef]
9. Ayling, J.; Chapman, A. Putting AI ethics to work: Are the tools fit for purpose? *AI Ethics* **2021**, *2*, 405–429. [CrossRef]
10. Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; Zhou, B. Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.* **2021**, *55*, 1–46. [CrossRef]
11. Morley, J.; Floridi, L.; Kinsey, L.; Elhalal, A. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In *Ethics, Governance, and Policies in Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 153–183. [CrossRef]
12. Hohma, E.; Boch, A.; Trauth, R.; Lütge, C. Investigating accountability for Artificial Intelligence through risk governance: A workshop-based exploratory study. *Front. Psychol.* **2023**, *14*, 1–17. [CrossRef] [PubMed]
13. Stix, C. Actionable principles for artificial intelligence policy: Three pathways. *Sci. Eng. Ethics* **2021**, *27*, 15. [CrossRef]

14. Dafoe, A. *AI Governance: A Research Agenda*; Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK, 2018; Volume 1442.
15. Miller, G.J. Stakeholder roles in artificial intelligence projects. *Proj. Leadersh. Soc.* **2022**, *3*, 100068. [CrossRef]
16. Wieringa, M. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 20), New York, NY, USA, 27–30 January 2020; pp. 1–18.
17. De Silva, D.; Alahakoon, D. An artificial intelligence life cycle: From conception to production. *Patterns* **2022**, *3*, 100489. [CrossRef]
18. Haakman, M.; Cruz, L.; Huijgens, H.; van Deursen, A. AI lifecycle models need to be revised. *Empir. Softw. Eng.* **2021**, *26*, 95. [CrossRef]
19. Suresh, H.; Guttag, J. A framework for understanding sources of harm throughout the machine learning life cycle. In Proceedings of the Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), New York, NY, USA, 5–9 October 2021; pp. 1–9. [CrossRef]
20. de Souza Nascimento, E.; Ahmed, I.; Oliveira, E.; Palheta, M.P.; Steinmacher, I.; Conte, T. Understanding development process of machine learning systems: Challenges and solutions. In Proceedings of the 2019 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), Porto de Galinhas, Brazil, 19–20 September 2019; pp. 1–6. [CrossRef]
21. Rybalko, D.; Portilla, I.; Kozhaya, J.; Ishizaki, K.; Hall, K.; Madan, N. *AI Model Lifecycle Management: What is ModelOps? A Technical Perspective*; IBM Point of View: Armonk, NY, USA, 2020.
22. Paulus, S.; Mohammadi, N.G.; Weyer, T. Trustworthy software development. In Proceedings of the Communications and Multimedia Security: 14th IFIP TC 6/TC 11 International Conference, CMS 2013, Magdeburg, Germany, 25–26 September 2013. pp. 233–247.
23. Yang, Y.; Wang, Q.; Li, M. Process trustworthiness as a capability indicator for measuring and improving software trustworthiness. In Proceedings of the Trustworthy Software Development Processes: International Conference on Software Process, ICSP 2009, Vancouver, BC, Canada, 16–17 May 2009; pp. 389–401.
24. Safonov, V.O. *Using Aspect-Oriented Programming for Trustworthy Software Development*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
25. Systems and Software Engineering—System Life Cycle Processes, ISO/IEC/IEEE. 2015. Available online: https://www.iso.org/standard/81702.html (accessed on 29 September 2023).
26. Developing Cyber-Resilient Systems: A Systems Security Engineering Approach, NIST. 2021. Available online: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-160v2r1.pdf (accessed on 29 September 2023).
27. Shiang-Jiun, C.; Yu-Chun, P.; Yi-Wei, M.; Cheng-Mou, C.; Chi-Chin, T. Trustworthy Software Development—Practical view of security processes through MVP methodology. In Proceedings of the 2022 24th International Conference on Advanced Communication Technology (ICACT), Pyeongchang, Republic of Korea, 13–16 February 2022; pp. 412–416.
28. IEEE Standards Association. *Addressing Ethical Concerns During Systems Design*; IEEE Standards Association: Piscataway, NJ, USA, 2021; Volume 7000.
29. IEEE Standards Association. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*; IEEE Standards Association: Piscataway, NJ, USA, 2019.
30. Weller, A. Transparency: Motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 23–40.
31. Wanner, J.; Herm, L.-V.; Heinrich, K.; Janiesch, C. The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electron. Mark.* **2022**, *32*, 2079–2102. [CrossRef]
32. Tyler, T.R. Why Do People Rely on Others? Social Identity and Social Aspects of Trust. In *Trust in Society*; Cook, K.S., Ed.; Russell Sage Foundation: New York, NY, USA, 2001; pp. 285–306.
33. Mökander, J.; Floridi, L. Operationalising AI governance through ethics-based auditing: An industry case study. *AI Ethics* **2023**, *3*, 451–468. [CrossRef] [PubMed]
34. Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H.; Yang, J.; Toner, H.; Fong, R. Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *arXiv* **2020**, arXiv:2004.07213.
35. Raji, I.D.; Smart, A.; White, R.N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; Barnes, P. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 20), New York, NY, USA, 27–30 January 2020; pp. 33–44.
36. NIST. *AI Risk Management Framework: Initial Draft*; NIST: Gaithersburg, MD, USA, 2022.
37. ISO. ISO/IEC JTC 1/SC 42 Artificial Intelligence. Available online: https://www.iso.org/committee/6794475/x/catalogue/p/1/u/0/w/0/d/0 (accessed on 19 December 2022).
38. ISO/IEC. *Information Technology—Artificial Intelligence—Overview of Trustworthiness in Artificial Intelligence*; ISO: Geneva, Switzerland, 2020. Available online: https://www.iso.org/standard/77608.html (accessed on 29 September 2023).
39. ISO/IEC. *Information Technology—Artificial Intelligence—Guidance on Risk Management*; ISO: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/77304.html (accessed on 29 September 2023).
40. Vakkuri, V.; Kemell, K.-K.; Kultanen, J.; Abrahamsson, P. The current state of industrial practice in artificial intelligence ethics. *IEEE Softw.* **2020**, *37*, 50–57. [CrossRef]

41. Burr, C.; Leslie, D. Ethical assurance: A practical approach to the responsible design, development, and deployment of data-driven technologies. *AI Ethics* **2023**, *3*, 73–98. [CrossRef]
42. Ashmore, R.; Calinescu, R.; Paterson, C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 111. [CrossRef]
43. AI Assurance Guide. Available online: https://cdeiuk.github.io/ai-assurance-guide/ (accessed on 21 September 2023).
44. Ada Lovelace Institute. *NMIP Algorithmic Impact Assessment User Guide*; Ada Lovelace Institute: Londond, UK, 2022.
45. High-Level Expert Group on Artificial Intelligence (AI HLEG). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment*; European Commission: Brussels, Belgium, 2020.
46. Vakkuri, V.; Kemell, K.-K.; Kultanen, J.; Siponen, M.; Abrahamsson, P. Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study. *arXiv* **2019**, arXiv:1906.07946.
47. Greenstein, B.; Rao, A. PwC 2022 AI Business Survey. Available online: https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-business-survey.html (accessed on 29 September 2023).
48. Wong, G.; Greenhalgh, T.; Westhorp, G.; Buckingham, J.; Pawson, R. RAMESES publication standards: Meta-narrative reviews. *J. Adv. Nurs.* **2013**, *69*, 987–1004. [CrossRef]
49. Snyder, H. Literature review as a research methodology: An overview and guidelines. *J. Bus. Res.* **2019**, *104*, 333–339. [CrossRef]
50. Aiethicist.org. Artificial Intelligence Resources; Aiethicist.org: 2022. Available online: https://www.aiethicist.org (accessed on 21 September 2023).
51. OECD. *Recommendation of the Council on Artificial Intelligence*; OECD/LEGAL/0449; OECD: Paris, France, 2019.
52. UNESCO. *Recommendation on the Ethics of Artificial Intelligence*; UNESCO: Paris, France, 2021.
53. US Federal Trade Commission (FTC). *Aiming for Truth, Fairness, and Equity in Your Company's Use of AI*; US Federal Trade Commission (FTC): Washington, DC, USA, 2021.
54. CEN-CENELEC Focus Group. *Road Map on Artificial Intelligence (AI)*; CEN-CENELEC: Brussels, Belgium, 2020.
55. Elam, M.; Reich, R. *Stanford HAI Artificial Intelligence Bill of Rights: A White Paper for Standford's Institute for Human-Centered Artificial Intelligence*; Stanford Human-Centered Artificial Intelligence: Stanford, CA, USA, 2022.
56. Felländer, A.; Rebane, J.; Larsson, S.; Wiggberg, M.; Heintz, F. Achieving a Data-driven Risk Assessment Methodology for Ethical AI. *Digit. Soc.* **2021**, *1*, 1–13. [CrossRef]
57. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef]
58. Reisman, D.; Schultz, J.; Crawford, K.; Whittaker, M. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability. *AI Now* **2018**. Available online: https://www.nist.gov/system/files/documents/2021/10/04/aiareport2018.pdf (accessed on 21 September 2023).
59. The Responsible Machine Learning Principles: A Practical Framework to Develop AI Responsibl. The Institute for Ethical AI & Machine Learning: London, UK. Available online: https://ethical.institute/index.html (accessed on 21 September 2023).
60. Loi, M.; Matzener, A.; Muller, A.; Spielkamp, M. *Automated Decision-Making Systems in the Public Sector: An Impact Assessment Tool for Public Authorities*; AW AlgorithmWatch gGmbH: Berlin, Germany, 2021.
61. The Public Voice. *Universal Guidelines for Artificial Intelligence*; The Public Voice: Burlington, VT, USA, 2018.
62. European Union. Summaries of EU Legislation; European Union: 2022. Available online: https://eur-lex.europa.eu/browse/summaries.html (accessed on 21 September 2023).
63. Braun, V.; Clarke, V. *Thematic Analysis*; American Psychological Association: Worcester, MA, USA, 2012.
64. Vaismoradi, M.; Snelgrove, S. Theme in qualitative content analysis and thematic analysis. *Forum Qual. Sozialforschung/Forum: Qual. Soc. Res.* **2019**, *20*. [CrossRef]
65. Mäntymäki, M.; Minkkinen, M.; Birkstedt, T.; Viljanen, M. Defining organizational AI governance. *AI Ethics* **2022**, *2*, 603–609. [CrossRef]
66. *Regulation 2021/0106*; Proposal for a Regulation of the Euopean Parliament and of the Council: Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. European Union: Brussels, Belgium, 2021.
67. *Regulation 2016/679*; Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). European Union: Brussels, Belgium, 2016.
68. *Regulation 2017/0003*; Proposal for a Regulation of the European Parliament and of the Council Concerning the Respect for Private Life and the Protection of Personal Data in Electronic Communications and Repealing Directive 2002/58/EC (Regulation on Privacy and Electronic Communications). European Union: Brussels, Belgium, 2017.
69. *Regulation 2022/0047*; Proposal for a Regulation of the European Parliament and of the Council on Harmonised Rules on Fair Access to and Use of Data (Data Act). European Union: Brussels, Belgium, 2022.
70. *Regulation 2020/0340*; Proposal for a Regulation of the European Parliament and of the Council on European Data Governance (Data Governance Act). European Union: Brussels, Belgium, 2020.
71. *Regulation 2021/0170*; Proposal for a Regulation of the European Parliament and of the Council on General Product Safety, Amending Regulation (EU) No 1025/2012 of the European Parliament and of the Council, and Repealing Council Directive

87/357/EEC and Directive 2001/95/EC of the European Parliament and of the Council. European Union: Brussels, Belgium, 2021.
72. *Directive 2022/0302*; Proposal for a Directive of the European Parliament and of the Council on Liability for Defective Products. European Union: Brussels, Belgium, 2022.
73. Almada, M.; Petit, N. The EU AI Act: Between Product Safety and Fundamental Rights. *Available SSRN* **2022**. [CrossRef]
74. *Directive 2022/0303*; Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive). European Union: Brussels, Belgium, 2022.
75. European Commission. The Digital Services Act Package. Available online: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package (accessed on 29 September 2023).
76. *Regulation (EU) 2022/2065*; European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and Amending Directive 2000/31/EC (Digital Services Act) . European Union: Brussels, Belgium, 2022.
77. *Regulation 2020/0374*; Proposal for a Regulation of the European Parliament and of the Council on Contestable and Fair Markets in the Digital Sector (Digital Markets Act). European Union: Brussels, Belgium, 2020.
78. *Directive 2005/29/EC*; European Parliament and of the Council of 11 May 2005 Concerning Unfair Business-to-Consumer Commercial Practices in the Internal Market and Amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive'). European Union: Brussels, Belgium, 2005.
79. *Charter 2000/C 364/01*; Charter of Fundamental Rights of the European Union (2000/C 364/01). European Union: Brussels, Belgium, 2000.
80. Kriebitz, A.; Lütge, C. Artificial intelligence and human rights: A business ethical assessment. *Bus. Hum. Rights J.* **2020**, *5*, 84–104. [CrossRef]
81. Seppälä, A.; Birkstedt, T.; Mäntymäki, M. From ethical AI principles to governed AI. In Proceedings of the 42nd International Conference on Information Systems (ICIS2021), Austin, TX, USA, 12–15 December 2021; pp. 1–17.
82. Hohma, E.; Boch, A.; Trauth, R. *Towards an Accountability Framework for Artificial Intelligence Systems*; TUM IEAI Whitepaper; TUM Institute for Ethics in Artificial Intelligence: Munich, Germany, 2022.
83. Anagnostou, M.; Karvounidou, O.; Katritzidaki, C.; Kechagia, C.; Melidou, K.; Mpeza, E.; Konstantinidis, I.; Kapantai, E.; Berberidis, C.; Magnisalis, I. Characteristics and challenges in the industries towards responsible AI: A systematic literature review. *Ethics Inf. Technol.* **2022**, *24*, 37. [CrossRef]
84. Gefen, D. E-commerce: The role of familiarity and trust. *Omega* **2000**, *28*, 725–737. [CrossRef]
85. Stern, M.J.; Coleman, K.J. The multidimensionality of trust: Applications in collaborative natural resource management. *Soc. Nat. Resour.* **2015**, *28*, 117–132. [CrossRef]
86. Hohma, E.; Burnell, R.; Corrigan, C.C.; Luetge, C. Individuality and fairness in public health surveillance technology: A survey of user perceptions in contact tracing apps. *IEEE Trans. Technol. Soc.* **2022**, *3*, 300–306. [CrossRef]
87. Burton, J.W.; Stein, M.K.; Jensen, T.B. A systematic review of algorithm aversion in augmented decision making. *J. Behav. Decis. Mak.* **2020**, *33*, 220–239. [CrossRef]