ROBONOMICS
The Journal of the Automated Economy
https://journal.robonomics.science

TAII Framework for Trustworthy AI Systems

Josef Baker-Brunnbauer [1] ✉

Abstract
Organisations and companies need practical tools and guidelines to kick-off the implementation of Trustworthy Artificial Intelligence (TAI) systems. AI development companies are still in the beginning of this process or have not even started yet. The findings of this article address to decrease the entry level barrier for AI ethics implementation by introducing the Trustworthy Artificial Intelligence Implementation (TAII) Framework. The outcome is comparatively unique given that it considers a meta perspective of implementing TAI within organisations. As such, this research aims to fill a literature gap for management guidance to tackle trustworthy AI implementation while considering ethical dependencies within the company. The TAII Framework takes a holistic approach to identify the systemic relationships of ethics for the company ecosystem and considers corporate values, business models, and common good aspects like the Sustainable Development Goals and the Universal Declaration of Human Rights. The TAII Framework creates guidance to initiate the implementation of AI ethics in organisations without requiring a deep background in philosophy and considers the social impacts outside of a software and data engineering setting. Depending on the legal regulation or area of application, the TAII Framework can be adapted and used with different regulations and ethical principles.

Keywords: Artificial Intelligence; Ethics; Trustworthiness; Business Model; Social Impact; Common Good

Type: Research paper

Citation: Baker-Brunnbauer, J. (2021). TAII Framework for Trustworthy AI Systems. *ROBONOMICS: The Journal of the Automated Economy*, 2, 17

[1] SocialTechLab.eu, Krumbachweg 5, 4060 Leonding, Austria; ORCID: 0000-0001-6805-8290; email: josef.baker-brunnbauer@socialtechlab.eu
✉ Corresponding author

1. Introduction

The evolution of AI systems can be seen from two different perspectives: as a negative strategy trying to prevent disasters and keep the AI system fulfilling its originally defined purpose or as a positive strategy that enriches the AI system's benefits to humanity (Boddington, 2021). The creation and usage of data for developing AI systems in an ethical manner evolves the need for data regulations in a connected digital world. The implementation needs to be done carefully in the area of tension between technology innovation and protection of privacy (Wachter, 2019). AI systems should not undermine the aims of the European General Data Protection Regulation (GDPR) (European Commission, 2021a). Instead, the GDPR should be seen as an enabler for implementing trustworthy AI systems. To govern AI systems can be a challenging task in the field of tension between different internal and external stakeholders, competition, and markets. Regulations should protect the fundamental rights of humans, but they can also generate tension with regard to international competition and innovation. Therefore, a successful set of AI guidelines requires an equal balance between technology and innovation, politics and state, economy and market, and humans and society with their environment (World Economic Forum, 2019).

The development of Trustworthy Artificial Intelligence (TAI) systems creates the need of practical tools and guidelines to initiate the implementation of AI ethics within companies for their products and services. This article introduces the Trustworthy Artificial Intelligence Implementation (TAII) Framework to tackle this need. As such, this research aims to fill a literature gap for management guidance to initiate trustworthy AI implementation while analysing ethical inconsistencies and dependencies for the planned AI system. Whereas other research on trustworthy AI (see section 2) has primarily focused on the definition and implementation of ethical principles, the TAII Framework is comparatively unique given that it considers the holistic perspective of developing and implementing trustworthy AI systems within organisations. Instead of starting directly with the implementation of ethical principles for the development of AI systems, the TAII Framework offers a management guidance to initiate the trustworthy AI implementation by starting with the analysis of ethical inconsistencies and dependencies for their planned AI system. The TAII Framework provides guidance for the involved stakeholders and considers these dependencies: corporate values, business models, and common good. In order to do this, this article is structured as follows. Section 2 describes the background to this work, section 3 explains the trustworthy AI approach of the European Commission, section 4 presents the TAII Framework, section 5 describes the practical transfer, and finally section 6 concludes the research. Trustworthy AI can generate major improvements in the areas of humans and society, private and public sectors, research and academia, availability of data and infrastructure, skills and education, governance and regulation, and funding and investment (European Commission AI HLEG, 2019a).

2. Background

An AI system should be seen as a socio-technical system whose impact is not only based on its design. Instead, the system should consider its broader environment, including purpose, training data, functionality and accuracy, scale of deployment, and the broader organisational, societal and legal context (Council of Europe, 2020). Data science opens up new ethical challenges in different research areas: the ethics of data, the ethics of algorithms, and the ethics of practices (Floridi & Taddeo, 2016). To prevent ethical inconsistencies during and after the development of AI systems, the implementation of AI ethics should be accomplished with the support of a multidisciplinary meta perspective by the key stakeholders. Starting with the translation of existing ethical principles does not lead to a common good solution for all as it focuses on already designated areas. For example, the use of an AI system to optimise animal farming may save costs and increase output but should also question systemic relationships to animal rights, diseases, environmental aspects, and dignity. AI ethics is a part of the ethical operation of a company. The existence of ethical guidelines is not a guarantee of utilisation and strongly grounded principles require legal mechanisms for implementation (Hagendorff, 2020). AI regulation approaches that take the social contract into account are ranked among the most open ones to interact with society in coproduction with the government (Delipetrev et al., 2020). Research to address the multidisciplinary

topic of AI ethics to embed political and societal context (Delipetrev et al., 2020) confirms that the answer of ethics implications of AI technologies requires a mix of law, design, and education (Calo, 2011). Besides making laws, the importance of discussions with society and academia to gain constant feedback is highlighted (Delipetrev et al., 2020).

More than 80 AI ethics initiatives published ethical principles and guidelines for AI system development and deployment (Hagendorff, 2020; Mittelstadt, 2019; Floridi et al., 2018; Twomey & Martin, 2020; KI Strategie Deutschland, 2020; European Union Agency for Fundamental Rights, 2020; Hickok, 2021; Cihon et al., 2020; Ryan & Stahl, 2021; Thiebes et al., 2021). Many initiatives envision to translate ethical high-level principles and abstract requirements, for example: fairness, transparency, or accountability, into mid- or low-level design requirements (Mittelstadt, 2019). The development of AI systems does not have empirically proven methods for translating principles into practical implementation (Mittelstadt, 2019). Needs and norms cannot be derived directly from mid-level design requirements without accounting for technology, context, application, and local norm elements. This requires normative decisions and the identification of coherences between principles, norms, and facts at each stage of the translation (Mittelstadt, 2019). Therefore, the AI ethics implementation has some challenges ahead, and a common alignment of some high-level principles is only a first small step as shared principles are no guarantee for a trustworthy AI system implementation. Most of the tools and methods for implementing ethical principles lack usability and do not provide enough practical support (Morley et al., 2019; Vakkuri et al., 2019). To implement AI ethics with a top-down approach (from general legal regulations to AI system developers) is more difficult than a bottom-up approach that starts with the requirements and settings within specific use-cases and applications (Mittelstadt, 2019). Empirical multidisciplinary bottom-up research might increase the speed of AI ethics implementation as it focuses directly on the needs and challenges of AI system developers. Besides giving attention to AI ethics on the development and deployment level, companies need to broaden their focus to the organisational ethics perspective. As AI engineers and developers will be constrained by their employers, AI ethics need to be aligned on the top levels of organisation. Research shows a big translation, implementation, and accountability gap between practical transfer and guidelines of ethical principles for AI system developers (Shklovski et al., 2021; Baker-Brunnbauer, 2021). This requires either additional skills for engineers or additional resources such as an 'ethics-officer-in-charge'. The TAII Framework supports the management of AI system developing companies to take actively the above-mentioned issues into account.

The implementation of AI ethics differs depending on the used technology, context and risk level of AI systems. The European Commission proposed six requirements for high-risk AI systems: clear liability and safety rules, information on the nature and purpose of an AI system, robustness and accuracy of AI systems, human oversight, quality of training datasets, and the keeping of records and data (European Commission, 2020a). In 2021, the European Commission released the regulatory framework proposal on AI (European Commission, 2021b) that classifies AI applications into four risk levels: minimal, limited, high, and unacceptable risk. High-risk AI systems will need to fulfil more requirements than others by undergoing a conformity assessment to reach the registration in the European Union database and to achieve the conformity declaration and CE marking (European Commission, 2021c). The TAII Framework supports the management of companies to develop trustworthy AI systems within each risk level.

3. Trustworthy AI

The European approach to trustworthy AI covers an ecosystem of excellence along the value chain from innovation and research to creating acceleration funding. Depending on the risk classification, AI systems that are developed or deployed within the European Union will need to fulfil the upcoming regulations (European Commission, 2021b). Therefore, the TAII Framework orientates on the TAI approach of the European Commission but can be adapted and used with different regulations and ethical principles. Furthermore, trustworthy AI will be shaped within an ecosystem of trust based on European fundamental rights and rules.

This will give users the trust and confidence to use AI systems (European Commission, 2020a). The implementation of trustworthy AI will gain public trust, clear responsibility, and enable 'dual advantage' (Floridi et al., 2018). Acceptance by the public and adoption of AI systems will be successful if the usage of AI technologies is seen as a low-risk and has meaningful areas of application (Floridi et al., 2018). The Sectoral Considerations on the Policy and Investment Recommendation for Trustworthy AI from the European Commission recommends a close collaboration between the industry and innovation ecosystems for research and transfer of trustworthy AI systems from ideation to rapid testing to deployment (European Commission AI HLEG, 2019b). Furthermore, it recommends proceeding within an open innovation culture within multidisciplinary research teams. The TAII Framework orientates on the European Commission's Ethics Guidelines for Trustworthy AI (European Commission AI HLEG, 2019c) but it is adaptable for others. The understanding of the European Commission is that trustworthy AI describes a human-centric and trustful development of AI systems to maximise the AI system benefits and minimise their risks (European Commission AI HLEG, 2019c). To generate a common understanding of the main terms within the organisation, the author recommends using a unique definition like the one from the AI High-Level Expert Group (European Commission AI HLEG, 2019d) what describes trustworthy AI with three components (lawful, ethical and robust) that should be aligned through the whole AI system's life cycle (European Commission AI HLEG, 2019c). The implementation of trustworthy AI should be an agile and continuous cycle during the whole AI system's life cycle and covers technical (architecture for trustworthy AI, ethics and rule of law by design, etc.) and non-technical (regulation, code of conduct, standardisation, certification, education and awareness to foster an ethical mind-set, stakeholder participation and social dialogue, diversity and inclusive design teams, etc.) methods (European Commission AI HLEG, 2019c).

4. TAII Framework

The development of trustworthy AI systems creates the need of practical tools and guidelines to initiate the implementation of AI ethics within companies for their products and services. Whereas other research on trustworthy AI (see section 2 and 3) has primarily focused on the definition and implementation of ethical principles, the Trustworthy Artificial Intelligence Implementation (TAII) Framework considers the holistic perspective of developing and implementing trustworthy AI systems. The TAII Framework has been developed by the author and offers a management guidance to initiate the trustworthy AI implementation by starting with the analysis of ethical inconsistencies and dependencies for their (planned) AI system along the value chain. The TAII Framework contains of twelve steps that will be continuously passed through during the whole AI system's life cycle.

The starting point for the implementation of trustworthy AI is the creation of an **AI system brief overview** (see Figure 1). This document describes the purpose, use case, and used input data of the AI system. The used and planned source data needs to be defined precisely as it is difficult to implement AI ethics when the training data consists of biased data (Brandon, 2021). The TAII Framework is iterative and refers to the whole AI system's life cycle. This has the benefit that the first iteration does not require too many details to initiate the AI ethics implementation. Within the following iterations, more details need to be added to clarify possible misunderstandings and to sharpen the picture. The defined and aligned **company values** will be used for the development of the company's **business model** as well as to build a solid ground for the AI system brief overview. Implementing AI ethics should correlate with a company culture that is based on ethics, morals, and values. To focus only on the implementation of AI ethics may cause contradictions with other company decisions. Therefore, it is recommended to see AI ethics as a part of the company values, morals, and adjacent ethics areas (Lauer, 2021). Otherwise, contradictions arise, such as implementing ethical principles for an AI system but buying raw material that was produced by disregarding fundamental human rights. Additionally, it is unclear how a company that acts unethically may (or based on regulations must) implement AI ethics without distorting the reality. Company management should consent, document, communicate, and transfer the values

to the company ecosystem (value proposition, stakeholder, supply chain, production etc.) and establish an internal ethics board that will lead the TAII.

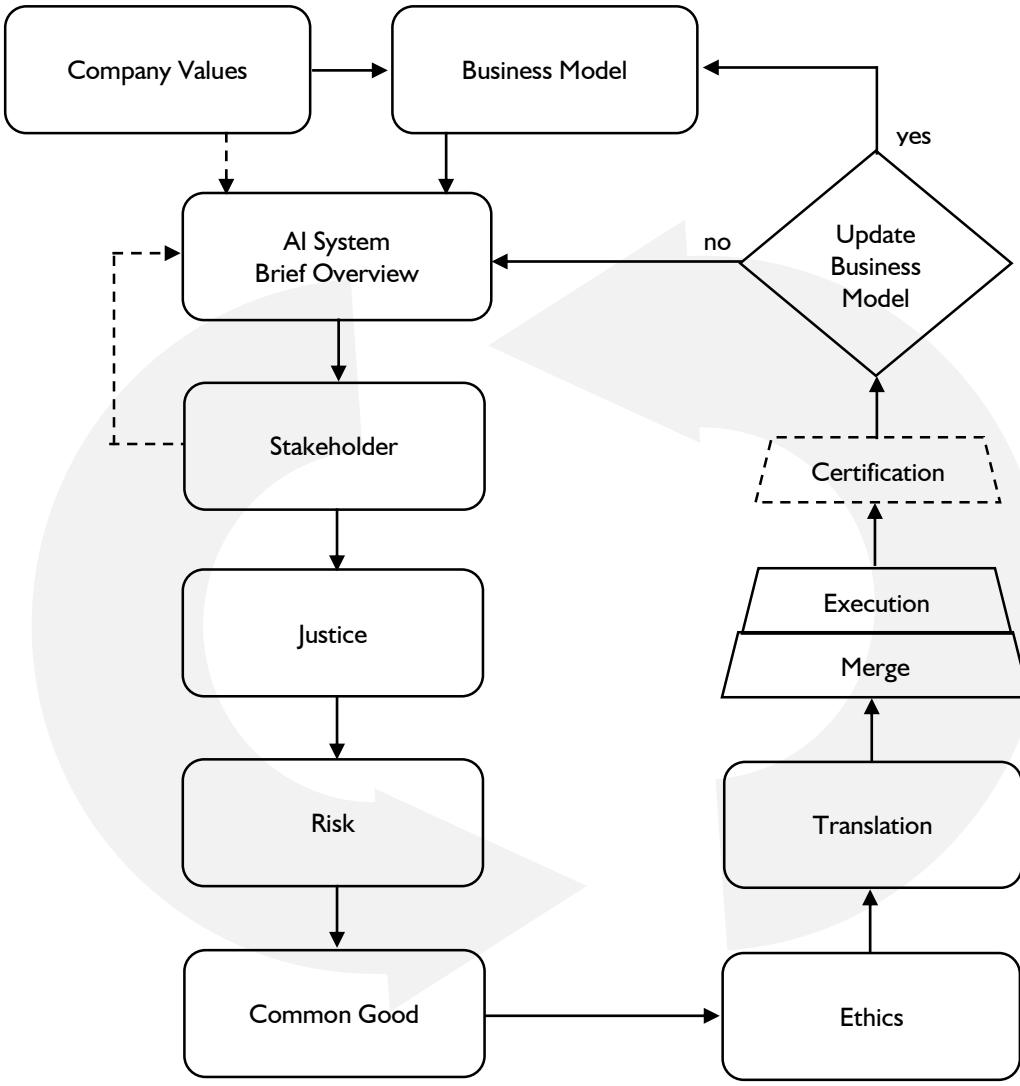

Figure 1. *Iteration of the Trustworthy AI Implementation (TAII) Framework for the AI system's life cycle*

To ensure business success, explore the AI system's ecosystem, and strengthen management's commitment to the TAII, a visualised **business model** should exist from the beginning (Baker-Brunnbauer, 2019). Facilitation by independent externals supports making decisions in situations where ethical principles and values are conflicting with each other. The next action is to define and to document the **stakeholders**. This lists all internally and externally involved people, groups, departments, companies, organisations, institutions, etc. and each should name a specific responsible persona. Stakeholders need to be categorised by role (contributor, responsible person, decision-maker, accountable person, supplier, developer, compliance person, deployer, user, governor, auditor, etc.). Next it is necessary to **justify** existing regulations and standards for the specific

AI system within the area of application. If those exist, the legal requirements and limitations need to be considered and implemented. Different proposals exist to apply governance models for technological development and law formulation processes (Delipetrev et al., 2020; European Parliamentary Research Service, 2020; DKE German Commission for Electrical, Electronic & Information Technologies of DIN and VDE, 2020; Ben-Israel et al., 2020) as well the Coordination Plan on Artificial Intelligence of the European Commission (European Commission, 2018) to safeguard a high standard of transparency, respect for democratic values, and legitimacy. It is the implementer's responsibility to be compliant with legal requirements and therefore this framework is not liable for the outcome.

The **risk** assessment of the ethical impact for the AI system is recommended and requires different measures depending on risk level. Existing legal requirements for the application area and different assessment methods (Ben-Israel et al., 2020; Mantelero, 2018; Artificial Intelligence Ethics Impact Group AIEI Group, 2020; Krafft & Zweig, 2019) can be chosen to classify risk and social impact. The risk assessment should consider the AI system's potential harm and the affected human groups based on the unintended results of the AI system. Ethical risk should not be unobservable or unquantifiable (European Parliamentary Research Service, 2020). The rating is often defined between a low and a maximum impact level. AI systems with a high impact level are called high-risk applications in literature (European Commission, 2020a; European Commission, 2021b; Council of Europe, 2020). Evaluating the risk in a two-dimensional matrix generates more clarity as the AI system may have different risk levels depending on the areas of application (Artificial Intelligence Ethics Impact Group AIEI Group, 2020). An industrial application may be classified as an ethical low risk level, but the AI system may strongly affect parameters like energy consumption or job replacement.

The systemic assessment of the AI system in compliance with the **common good** via the Sustainable Development Goals (SDGs) (United Nations Sustainable Development Goals, n.d.) and the Universal Declaration of Human Rights (UDHR) (United Nations Universal Declaration of Human Rights, n.d.) will analyse dependencies and patterns within the organisational ecosystem (Systemic Society, n.d.). The AI system definition from the Council of Europe (2020) that says the AI system should include its broader environment and its organisational, societal and legal context, offers a good basis to explore the AI system's broader impact on the SDGs and UDHR. During the assessment of the SDGs, organisations will identify to which goals their AI system will contribute in a positive or negative way. Depending on the application or used technology, some goals are possibly not fitting for the AI system. Those can be skipped but the evaluation of the AI system's impact on the SDGs should be assessed critically. To improve transparency, the documentation of all answers and future iterations of the TAII Framework is recommended, which also generates new input. The reflection of the AI system with the 17 SGDs will broaden perspective, demonstrate the systemic dependencies of resources that should be used in a sustainable way, and question the connections for possible negative interference regarding the UDHR. AI ethics is only a part of how the company and its products and services will contribute to the environment and humanity.

The next step generates the list of **ethical** requirements and guidelines. Many AI ethics guidelines are already developed (Fjeld et al., 2020; Hagendorff, 2020; Jobin et al., 2019) and the AI system developer needs to align which ones fit best based on different factors such as (inter-)national regulations, legal requirements, field of application, standardisation, etc. Within this research, the European Commission's approach to achieving trustworthy AI systems (European Commission, 2020a; European Commission AI HLEG, 2019c) has been chosen. This implies four principles and seven key requirements. The requirements are as follows: human agency and oversight; robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental well-being; accountability (European Commission, 2020a; European Commission AI HLEG, 2019b; European Commission AI HLEG, 2019c). The defined ethical principles and requirements need to be appropriately **translated** and transferred for the AI system ecosystem. This starts with the mapping of the previous defined principles with the application-related ecosystem. It is not

effective to use a one-size-fits-all approach since, for example, the requirement 'transparency' will have different aspects depending on the use-case, application, and context. Transparency can be understood in different ways, for example to publish a short statement about the AI system's algorithm (communication), to make the whole source code of the AI system's algorithm public (decision-making-process), or to achieve a certification that's state the fulfilment of some requirements for the AI system's algorithm (comply standards). For the first iterations it helps to follow checklist questions (European Commission AI HLEG, 2020) and to implement a specific mapping method in a more progressed iteration. This can be a method such as Values, Criteria, Indicators, Observables (VCIO) (Artificial Intelligence Ethics Impact Group AIEI Group, 2020), Value Sensitive Design (VSD) (Umbrello & van de Poel, 2021), Design for Values (Dignum, 2019), Value Sensitive Software Development (VSSD) (Aldewereld et al., 2015), data-driven research framework (DaRe4TAI) (Thiebes et al., 2021), applied ethical AI typology (Morley et al., 2019), Artificial Intelligence Regulation (AIR) (Hagendorff, 2020) etc. Companies will need to take both actions and self-responsibility to make the best-fitting decisions for the transfer of ethical principles.

After the mapping of the ethical principles, the **merge** of the previously assessed input factors (AI system brief overview, stakeholder, justice, risk, common good, ethics) starts. The goal of the merge is to define the current state, to visualise dependencies, and to plan the next tasks for continuous improvement. The **execution** verifies, tests, and implements the results. During the TAII all considerations and actions should be documented, and the responsibility of tasks must be clarified to achieve transparency. The execution of a mandatory (cf. AI high-risk applications (European Commission, 2021b)) or optional **certification** or safety assessment for the AI system will increase transparency and trust. Different institutions are working on certification standards (European Commission, 2020a; Council of Europe, 2020; Cremers et al., 2019; DKE German Commission for Electrical, Electronic & Information Technologies of DIN and VDE, 2020; Braband & Schäbe, 2020). As the AI system may change during every TAII iteration, the responsible stakeholder for the TAI implementation needs to reconsider if a new or re-certification (assessment) is required or necessary.

After passing through the last step 'certification' of the TAII Framework (see Figure 1) the **next iteration** starts with an update of the AI system brief overview or if necessary, the business model. Some parameters will be changed or updated within future iterations as product, service, market, and society are evolving. The iterations are endless as long the AI system's life cycle reaches its end. The area of application, risk assessment, maturity level of the AI system, and external factors (technology, market, society etc.) will influence the speed and frequency of the iteration interval that needs to be defined by the AI system developer or deployer. The TAII Framework and its twelve steps can be used for existing and for future AI systems during the system's whole life cycle. Nevertheless, the success of the TAII is based on the defined priority given to AI ethics, planned and allocatable resources, and the commitment of all core stakeholders to implement and regularly pass through the TAII Framework. To improve transparency, a documentation of all answers and future iterations of the TAII Framework is recommended by the author. The author proposes to apply the TAII Framework to all risk levels of AI systems, including low-risk classification with less restrictive actions. Skipping TAII for low-risk AI systems should not be considered as it may oversee broader social impacts and dependencies. AI ethics is a part of the ethical behaviour of the organisation and influences its contribution to the environment and society.

5. Implementation challenges of TAI

Besides following guidance, some challenges for AI system developers and deployers may arise. The implementation of unintentional negative consequences occurs when AI systems are deployed without compliance efforts and without a robust governance (Eitel-Porter, 2021). The TAII Framework helps companies to reduce those unintentional negative consequences by exploring the broader ecosystem and analysis of hidden dependencies. Companies that are more innovative are exposed to a higher risk of implementing unintentional negative consequences. Reasons for this are short development cycles, lack of technical understanding, no established quality assurance, usage of AI outside the original defined context, improper combination of data,

and unreported concerns of employees (Eitel-Porter, 2021). Ethical AI systems also require strong governance controls, including process management audit procedures. Implementing (AI) ethics generates costs for AI system development. Aligning the stakeholders, defining a responsible ethics team, considering interdisciplinary perspectives, adapting development and testing concepts, etc. requires additional resources. During the analyses of the TAII Framework stakeholder step, costs and resources can be planned, calculated and allocated. Ethical considerations may also conflict with commercial interest. The evaluation of the AI system's broader impact, including the ecosystem outside the technical development environment, by using the TAII Framework supports to identify possible conflicts in an early stage. Finally, the executive management needs to commit and to make the final decisions.

During the second European AI Alliance Assembly a survey identified the following challenges to AI system deployment: explainability, trust and accuracy, privacy, ethics, predictive accuracy, and transparency (European Commission, 2020b). Research shows that the majority of deployed AI systems are neither transparent nor comprehensible to their users. Rather, they are only interpretable to the engineers used to debug the algorithm (Bhatt et al., 2020). Using the TAII Framework explores the implications for common good and analyses affected stakeholders including the engineers but also targeting the broader audience of non-technicians. A successful approach is to extend and split transparency and explainability into two layers: a version on a technical level that is understandable for engineers and another one that is more abstract and understandable for the audience of non-technicians (Bhatt et al., 2020). Transparency exists in many different forms: data, decision-making process, communication, etc. and it limits harm and increases people's trust (Shklovski et al., 2021). Documentation is a key element for all twelve steps of the TAII Framework to achieve transparency and requires both a clear process and a routine 'when, what, who and how-to' protocol (Shklovski et al., 2021). Considering the complexity of AI systems, it is hardly possible to achieve total explainability and it is difficult to determine how much explainability is required (Shklovski et al., 2021). The analysis of the broader stakeholders (technical and non-technical) following the TAII Framework creates an understanding of the needs, expectations and concerns of the affected groups. Based on the stakeholder interviews, legal requirements, risk assessment and common good aspects, AI system developer and deployer need to decide about the level of explainability.

Values and ethics should be considered to be inseparable and intertwined. As all stakeholders should have a common understanding of AI and ethical terms, the author recommends using an aligned definition of specialist terms. Already this first step can be challenging for companies as there are no commonly agreed definitions. One approach is to follow the terms of the European Commission (European Commission AI HLEG, 2019d; European Commission AI HLEG, 2020). Some initiatives propose AI ethics implementation with a 'technical and design' expertise by providing technical concepts for privacy, fairness, etc. (Greene et al., 2019). Technical definitions or explanations for such technological proposals rarely exist (Hagendorff, 2020) and the results of practical implementations are unclear. The TAII Framework supports organisations to start exploring AI ethics from a broader non-technical perspective and makes it possible to deep dive into technical and design implementations in later iterations. AI ethics implementation will be difficult to solve only with a technical approach (Mittelstadt, 2019) and there is not a universal implementation formula and it cannot be realised by a blueprint. The TAII Framework offers a holistic guidance, and the concrete actions need to be aligned depending on the area of application, context and used AI technology. Accountability is a central topic for the development of AI systems and data processing, but it comes along with many unanswered questions about the responsibilities of AI system developers (Shklovski et al., 2021; Baker-Brunnbauer, 2021). As this is a general challenge for all AI system developer and deployer, a common legal framework that specifies the accountability is required. During the pass through of the TAII Framework all important actions that might influence the accountability should be documented as long as there is no binding legal framework. The risk assessment can be conducted with the two-axis Risk Matrix (Artificial Intelligence Ethics Impact Group AIEI Group, 2020). It defines the risk level on the x-axis by the intensity of possible harm (the number of people that are negatively impacted, the negative impact on society, and the impact on fundamental rights, equality, and social justice).

The y-axis visualises the dependency of the potentially affected stakeholders on the AI system's actions and decisions by considering control (lower demand for regulation as the AI system operates without human intervention), switchability (possibility of exchanging the AI system and monopolistic dependency to a supplier or producer) and redress (the time needed to understand and correct an unintended AI system outcome). AI systems with the same AI technology can be assigned to different risk levels. Therefore, the application context and area, purpose, training data, and the AI system requirements for the risk assessment needs to be included. AI high-risk applications will need to undergo an extra certification to gain the CE marking for distribution and usage within the European Union (European Commission, 2021b). The TAII Framework supports companies developing trustworthy AI systems before undergoing the conformity assessment.

Without existing legal regulations, the classification as well as the whole AI ethics implementation is not binding but helps companies prepare for upcoming regulatory implementation, managing their product portfolio, and improving their social impact (Baker-Brunnbauer, 2021). Engineers need to engage openly with internal and external stakeholders on how ethical principles can be implemented. Their organisation should support this engineer engagement, expand documentation practices, and should enable support and exchange with public authorities, organisations and stakeholders (Shklovski et al., 2021). The TAII Framework can be used in stakeholder workshops to introduce the participants in AI ethics, its challenges and to generate an overall understanding of the trustworthy AI implementation requirements. All stakeholders are responsible for sharing their opinions and for making specific proposals for accountable structures. Therefore, organisations need to encourage a culture of open learning and ensure the distribution of responsibility and accountability within the company by implementing standards, assessments, documentation, and testing (Shklovski et al., 2021). Besides the AI system's development life cycle, the Technology Readiness Level (TRL) of the AI technology itself needs to be assessed (Martínez-Plumed et al., 2020). Each of the ethical principles needs to be translated into design and technical requirements that reflects the principles' aim (Anabo et al., 2019; La Fors, 2019). This requires a translation in the level of abstraction from principles to micro ethics (Hagendorff, 2020) for reducing abstract norms and generating a Minimum Viable Ethical Product (MVEP) that is useful for people with diverse backgrounds (Jacobs & Huldtgren, 2021). AI system ideation and development workshops using the TAII Framework generate additional value of sustainability and common good along the value proposition and its technical feature set. The TAII Framework evolves its power by being used in an early stage of the AI system development to include the broader perspectives of involved stakeholders along the value chain, legal requirements, risk consequences and common good from the beginning. Non-transparent and incomprehensible AI systems will never be socially acceptable because humans will feel controlled by them (Floridi & Taddeo, 2016).

6. Conclusion

Talk about AI ethics is increasing, and management consulting companies view the topic as a future game changer for companies. Empirical research shows that the practical implementation of AI ethics is lagging, and that companies are in the early stage of seeking out guidance. The Trustworthy AI Implementation (TAII) Framework initiates this process and supports management in orienteering and implementing AI ethics within their organisation. With the TAII Framework the management team and its ethics board can explore the systemic dependencies inside and outside their organisation. TAII cannot change a purposefully unethical company strategy and it supports those whose intention is to take self-responsibility for the environment and its living beings. Management needs to generate awareness and implement ethical guidelines within their organisation. During the assessments the focus is on questions that cannot be easily answered. Diverse teams and knowledge plus independent consultancy support the implementation of TAI. The arising questions and outcomes must be aligned with applicable laws and regulations. The implementation of TAI requires a strong long-term company commitment to the development, deployment, and usage of trustworthy AI systems. Tensions between principles and values may arise between the assessment and implementation of trustworthy AI as there is no solution that fits for all AI systems. Stakeholders should analyse the ethical dilemmas with

evidence-based reflections and avoid making random decisions. Challenges may arise during the translation and implementation of the multidiscipline topic as values, morals, and ethics are not understandable for AI systems by default and commercial interest may generate tensions. A company's culture and its values as well their business model will strongly influence the success of trustworthy AI implementations.

References

Aldewereld, H., Dignum, V., & Tan, Y. (2015). Design for Values in Software Development. In van den Hoven, J., Vermaas, P., & van de Poel, I. (Eds.). *Handbook of Ethics, Values, and Technological Design*. Dordrecht, Netherlands: Springer, pp. 831-845. https://doi.org/10.1007/978-94-007-6970-0_26

Anabo, I.F., Elexpuru-Albizuri, I., & Villardón-Gallego, L. (2019). Revisiting the Belmont Report's ethical principles in internet-mediated research: perspectives from disciplinary associations in the social sciences. *Ethics and Information Technology, 21*, 137–149. https://doi.org/10.1007/s10676-018-9495-z

Artificial Intelligence Ethics Impact Group AIEI Group. (2020). *From Principles to Practice. An Interdisciplinary Framework to Operationalise AI Ethics*. Retrieved 9.1.2021 from https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf

Baker-Brunnbauer, J. (2019). *Business Model Innovation in a Paradoxical Area of Conflict* (Executive Summary). https://doi.org/10.13140/RG.2.2.24272.66566

Baker-Brunnbauer, J. (2021). Management perspective of ethics in artificial intelligence. *AI and Ethics, 1*, 173–181. https://doi.org/10.1007/s43681-020-00022-3

Ben-Israel, I., Cerdio, J., Ema, A., Friedman, L., Ienca, M., Mantelero, A., Matania, E., Muller, C., Shiroyama, H., & Vayena, E. (2020). *Towards Regulation of AI Systems*. Council of Europe Study. Retrieved 9.1.2021 from https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., & Eckersley P. (2020). *Explainable Machine Learning in Deployment*. Retrieved from https://arxiv.org/abs/1909.06342v4

Braband, J., & Schäbe, H. (2020). On safety assessment of artificial intelligence. *Dependability, 4*, 25-34. https://doi.org/10.21683/1729-2646-2020-20-4-25-34

Council of Europe. (2020). *Possible introduction of a mechanism for certifying artificial intelligence tools and services in the sphere of justice and the judiciary*. Retrieved 14.12.2020 from https://rm.coe.int/feasability-study-en-cepej-2020-15/1680a0adf4

Cremers, A. B., Englander, A., Gabriel, M., Hecker, D., Mock, M., Poretschkin, M., Rosenzweig, J., Rostalski, F., Sicking, J., Volmer, J., Voosholz, J., Angelika Voss, A., & Wrobel, S. (2019). *Trustworthy Use of Artificial Intelligence. Priorities from a philosophical, ethical, legal, and technological viewpoint as a basis for certification of artificial intelligence*. Retrieved 9.1.2021 from https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_Thrustworthy_AI.pdf

Boddington, P. (2021). AI and moral thinking: how can we live well with machines to enhance our moral agency? *AI and Ethics, 1*, 109-111. https://doi.org/10.1007/s43681-020-00017-0

Brandon, J. (2021). Using unethical data to build a more ethical world. *AI and Ethics, 1*, 101–108. https://doi.org/10.1007/s43681-020-00006-3

Calo, M.R. (2011). Peeping Hals. *Artificial Intelligence, 175(5-6)*, 940-941. https://doi.org/10.1016/j.artint.2010.11.025

Cihon, P., Maas, M.M. & Kemp, L. (2020). Fragmentation and the Future: Investigating Architectures for International AI Governance. *Global Policy, 11(5)*, 545-556. https://doi.org/10.1111/1758-5899.12890

Council of Europe. (2020). Ad hoc Committee on Artificial Intelligence (CAHAI). Feasibility Study. Retrieved 18.1.2021 from https://rm.coe.int/cahai-2020-23-final-eng-feasibility-study-/1680a0c6da

Delipetrev, B., Tsinaraki, C. & Kostic, U. (2020). *Historical Evolution of Artificial Intelligence*. EUR 30221 EN. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/801580

Dignum, V. (2019). *HUMANE AI - Toward AI Systems that Augment and Empower Humans by Understanding Us, our Society and the World Around Us*. Retrieved 9.1.2021 from https://www.humane-ai.eu/wp-content/uploads/2019/11/D13-HumaneAI-framework-report.pdf

DKE German Commission for Electrical, Electronic & Information Technologies of DIN and VDE. (2020). *German Standardization Roadmap on Artificial Intelligence*. Retrieved 9.1.2021 from https://www.dke.de/resource/blob/2017010/99bc6d952073ca88f52c0ae4a8c351a8/nr-ki-english---download-data.pdf

Eitel-Porter, R. (2021). Beyond the promise: implementing ethical AI. *AI and Ethics, 1*, 73–80. https://doi.org/10.1007/s43681-020-00011-6

European Commission. (2018). *Annex to the Coordinated Plan on Artificial Intelligence*. Retrieved 5.8.2020 from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56017

European Commission. (2020a). *White paper on artificial intelligence: a European approach to excellence and trust*. Retrieved 19.2.2020 from https://ec.europa.eu/info/publications/white -paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission. (2020b). *Second European AI Alliance Assembly*. Retrieved 8.1.2021 from https://ec.europa.eu/digital-single-market/en/news/second-european-ai-alliance-assembly

European Commission. (2021a). *Data protection. Rules for the protection of personal data inside and outside the EU*. Retrieved 24.2.2021 from https://ec.europa.eu/info/law/law-topic/data-protection_en

European Commission. (2021b). *Regulatory framework proposal on Artificial Intelligence*. Retrieved 8.7.2021 from https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

European Commission. (2021c). *Excellence and trust in artificial intelligence*. Retrieved 8.7.2021 from https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_en

European Commission AI HLEG. (2019a). *Policy and Investment Recommendations for Trustworthy Artificial Intelligence*. Retrieved 8.1.2021 from https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence

European Commission AI HLEG. (2019b). *Sectoral considerations on the policy and investment recommendations for trustworthy artificial intelligence*. Retrieved 8.1.2021 from https://futurium.ec.europa.eu/sites/default/files/2020-07/Sectoral%20Considerations%20On%20The%20Policy%20And%20Investment%20Recommendations%20For%20Trustworthy%20Artificial%20Intelligence_0.pdf

European Commission AI HLEG. (2019c). *Ethics Guidelines for Trustworthy AI*. Retrieved 8.1.2021 from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

European Commission AI HLEG. (2019d). *A Definition of AI: Main capabilities and disciplines*. Retrieved 8.1.2021 from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60651

European Commission AI HLEG. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Retrieved 8.1.2021 from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

European Parliamentary Research Service. (2020). *Artificial intelligence: From ethics to policy*. Retrieved 6.1.2021 from https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641507/EPRS_STU(2020)641507_EN.pdf

European Union Agency for Fundamental Rights. (2020). *AI policy initiatives (2016–2020)*. Retrieved 4.7.2020 from https://fra.europa.eu/en/project/2018/artificial-intelligence-big-data-and-fundamental-rights/ai-policy-initiatives

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI*. Berkman Klein Center Research Publication, 2020-1, 1-39. http://dx.doi.org/10.2139/ssrn.3518482

Floridi, L., & Taddeo, M. (2016). What is Data Ethics? *Philosophical Transactions of the Royal Society A, 374(2083)*, 20160360. http://dx.doi.org/10.1098/rsta.2016.0360

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences* (HICSS, 2019), pp. 2122–2131. https://doi.org/10.24251/HICSS.2019.258

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines, 30*, 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics, 1*, 41–47. https://doi.org/10.1007/s43681-020-00008-1

Jacobs, N., & Huldtgren, A. (2021). Why value sensitive design needs ethical commitments. *Ethics and Information Technology, 23*, 23-26. https://doi.org/10.1007/s10676-018-9467-3

Jobin, A., Ienca, M. & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1(9)*, 389–399. https://doi.org/10.1038/s42256-019-0088-2

KI Strategie Deutschland. (2020). *Artificial Intelligence Strategy of the German Federal Government*. Retrieved 2.3.2021 from https://www.ki-strategie-deutschland.de/files/downloads/Fortschreibung_KI-Strategie_engl.pdf

Krafft, T. D., & Zweig, K.A. (2019). *Transparenz und Nachvollziehbarkeit algorithmenbasierter Entscheidungsprozesse* | Ein Regulierungsvorschlag. Retrieved 17.2.2021 from https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf

La Fors, K., Custers, B. & Keymolen, E. (2019). Reassessing values for emerging big data technologies: integrating design-based and application-based approaches. *Ethics and Information Technology, 21*, 209–226. https://doi.org/10.1007/s10676-019-09503-4

Lauer, D. (2021). You cannot have AI ethics without ethics. *AI and Ethics, 1*, 21–25. https://doi.org/10.1007/s43681-020-00013-4

Mantelero, A. (2018). AI and Big Data: A blueprint for human rights, social and ethical impact assessment. *Computer Law & Security Review, 34(4)*, 754-772. https://doi.org/10.1016/j.clsr.2018.05.017

Martínez-Plumed, F., Gómez, E. & Hernández-Orallo, J. (2020). *AI Watch, Assessing Technology Readiness Levels for Artificial Intelligence*, Joint Research Center European Commission. https://doi.org/10.2760/15025

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*, 501–507. https://doi.org/10.1038/s42256-019-0114-4

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2019). *A Typology of AI Ethics Tools, Methods and Research to Translate Principles into Practices*. Retrieved 10.10.2020 from https://aiforsocialgood.github.io/neurips2019/accepted/track2/pdfs/26_aisg_neurips2019.pdf

Ryan, M., & Stahl, B.C. (2021). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society, 19(1)*, 61-86. https://doi.org/10.1108/JICES-12-2019-0138

Shklovski, I., ANE, Data Ethics ThinkDoTank, & IEEE. (2021). *Addressing Ethical Dilemmas in AI: Listening to Engineers*. Retrieved 27.1.2021 from https://futurium.ec.europa.eu/sites/default/files/2021-01/Addressing%20Ethical%20Dilemmas%20in%20AI%20%E2%80%93%20Listening%20to%20the%20Engineers.pdf

Systemic Society. Deutscher Verband für systemische Forschung, Therapie, Supervision und Beratung e.V. (n.d.). *Systemische Methoden*. Retrieved 21.8.2021 from https://systemische-gesellschaft.de/systemischer-ansatz/methoden

Thiebes, S., Lins, S. & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets, 31*, 447-464. https://doi.org/10.1007/s12525-020-00441-4

Twomey, P., & Martin, K. (2020). *A Step to Implementing the G20 Principles on Artificial Intelligence: Ensuring Data Aggregators and AI Firms Operate in the Interests of Data Subjects*. Retrieved 8.1.2021 from https://www.g20-insights.org/wp-content/uploads/2020/04/g20-principles-artificial-intelligence-data-aggregators-ai-firms-1586167851.pdf

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00038-3

United Nations Sustainable Development Goals. (n.d.). Retrieved 9.1.2021 from https://sdgs.un.org

United Nations Universal Declaration of Human Rights. (n.d.). Retrieved 9.1.2021 from https://www.un.org/en/universal-declaration-human-rights

Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M., & Abrahamsson. P. (2019). *Ethically aligned design of autonomous systems: Industry viewpoint and an empirical study*. Retrieved from http://arxiv.org/abs/1906.07946

Wachter, S. (2019). Data protection in the age of big data. *Nature Electronics*, *2*, 6-7. https://doi.org/10.1038/s41928-018-0193-y

World Economic Forum. (2019). *AI Governance: A Holistic Approach to Implement Ethics into AI*. Retrieved 28.4.2020 from https://www.weforum.org/whitepapers/ai-governance-a-holistic-approach-to-implement-ethics-into-ai

Received: 27/07/2021
Revised: 23/08/2021
Revised: 29/08/2021
Accepted: 30/08/2021