The document reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

InSecTT:

Intelligent Secure Trustable Things

The Development of Ethical and Trustworthy AI Systems Requires Appropriate Human-Systems Integration: A White Paper

Document Type	Whitepaper
Primary Author(s)	Peter Moertl, Nikolai Ebinger | ViF
Document Version / Status	0.2 | Draft
Distribution Level	PU (public)

Project Acronym	InSecTT
Project Title	Intelligent Secure Trustable Things
Project Website	https://www.insectt.eu/
Project Coordinator	Michael Karner | VIF | michael.karner@v2c2.at
JU Grant Agreement Number	876038
Date of latest version of Annex I against which the assessment will be made	2021-06-25

InSecTT has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876038. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Austria, Sweden, Spain, Italy, France, Portugal, Ireland, Finland, Slovenia, Poland, Netherlands, Turkey

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Executive Summary

The development of artificial intelligence (AI) technologies experiences worldwide an ongoing challenge to become trustworthy and ethical for users and the general public. This challenge currently stands between the promise of AI to create immense societal and individual impact and its realization. Because of this, possible large marketplaces still remain hesitant or closed. We have investigated this problem and identified potential solutions in the InSecTT project, a large international EU research and development project that investigates ethical smart technologies[1]. Thereby, working with industrial and research partners, we assert that developing trustworthy AI technologies is not foremost a technical challenge but increasingly an organizational and structural challenge that results from applying traditional ways of conceiving, designing, and selling technologies to new types of problems. Because AI technologies can shift the role that humans and society see as acceptable, traditional development processes that rely on strict separation of specialties are overburdened. In our view a research and development approach for trustworthy AI systems should put human concerns and needs at the center of the development process to effectively integrate humans and systems. Also, such approach should be based on EU guidelines for developing ethical AI. Our proposed approach to develop trustworthy AI systems combines these two directions and centers around the assessment of trustworthiness risks through intensive user involvement prior to the elicitation of system requirement that are then managed throughout the system's life-cycle. Also, the approach includes concrete recommendations to establish the organizational prerequisites that would enable organizations to implement the design process recommendations.

In this white paper, we describe the Human-Systems Integration (HSI) approach and motivate the underlying principles in some detail. The white paper intends to inform managers of technical organizations and product managers as well as principal investigators to set up the prerequisites for trustworthy AI. The white paper also wants to solicit inputs for further refinement and discussion.

Keywords: Trustworthy AI, Human-Centered Process, Ethical AI

[1] https://www.insectt.eu/

2 What is the Problem?

The market for smart technologies that utilize artificial intelligence (AI) is growing at high rates, but uptake of these technologies is lagging behind due to increasing lack of user trust and acceptance of these technologies. Among many examples, in 2020, Amazon had to stop selling its face recognition software due to racially biased categorizations and Microsoft soon followed suit[2]. Similarly, warned the German Bundesnetzagentur against the use of certain intelligent toys because of potential spying on customers[3]. There are many examples that have triggered debates on how to create ethical AI systems that are consistent with the interests and rights of its users. These examples emphasize the drawbacks of smart technologies that result from their often unintended consequences when users interact with them in the real world.

Smart technologies often shift the roles and responsibilities of those surrounding them and these impacts are often difficult to see beforehand. A single police agent with biased racial preferences is one thing but an automated facial recognition system with such biased views shifts the problem to another level as hundreds of thousands of biased categorizations could occur.

Thereby we define a trustworthy smart system as one that facilitates its user's trust as "… the attitude that an agent [smart system] will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [1]. Trustworthy systems must be lawful, ethical, and robust [2], which makes ethical compliance a necessary but not sufficient precondition for trustworthiness.

Based on the recently proposed EU AI act, we consider an ethical system as one that honors the rights of the human user and thereby conforms to moral principles, specifically ensuring the autonomy of humans to decide for themselves, not exploit their data and rights for other purposes, and ultimately serve the human, rather than purely those selling products[4].

"Smart" functions increasingly take over the cognitive and manual tasks that previously only humans could perform. Smart systems recognize preferences, understand voice commands, keep the distance to the vehicle driving ahead, or provide diagnostic information to medical doctors. Thereby in most cases beyond the most repetitive and simple environments, do smart systems usually not replace the human operator but assist them in their work and daily life. Full automation is often too expensive and not realistic, given liability concerns. A survey among 500 decision-makers from German companies shows that smart AI systems are mostly intended to assist human operators but not to replace them [4]. Whereas fully automated systems are often implied by public debates about the capabilities of artificial intelligence, the limits are often not clearly stated and fiction and reality are blurred. Thereby, AI developing organizations are motivated to overstate the smart technologies they produce to remain relevant. This can result in misunderstandings about the actual ability of AI based technology and produce societal backlashes against the technology without understanding the real use cases. Because smart systems cannot yet decide about their contextual enabling and

[2] https://www.reuters.com/technology/exclusive-amazon-extends-moratorium-police-use-facial-recognition-software-2021-05-18/

[3] https://www.heise.de/news/Smart-Toys-Bundesnetzagentur-warnt-vor-Spionage-Spielzeug-6300179.html

[4] The term ethical is commonly defined as conforming to "the moral principles that govern a person's behaviour or how an activity is conducted" [3]. This definition is rather abstract and to create an helpful understanding for this context, we select, based on the EU guidelines for trustworthy AI, the moral principles of "autonomy", "privacy", and "human well-being and diversity" as most critical enablers of ethics. Examples may be useful to exemplify this: A system that changes the human role from an active, decision making role into a passive, reactive role is, according to this definition, not ethical. A system that exploits sensitive data from human users for other purposes than those to whom they belong, is not ethical. Finally, a system that only serves a select few and discriminates against others, is also not ethical. In this way, a system that is ethical may be trusted, therefore, it is trustworthy.

disabling, solve untrained situations, or take responsibility for failures, realistic use cases involve the human operator to decide in these situations. This is the case in many different operational domains such as medicine or security; a medical diagnostic system can help the medical doctor during the diagnosis process but not take responsibility for medical decision making; an effective security system can smartly detect anomalies within large data sets and then inform the operator with the needed information to resolve a security breach. These represent new roles for the human and the smart technology needs to be designed for such teamwork.

The main challenge of teaming AI with humans is that humans who are for some periods of time out-of-the-loop, are difficult to bring back into the loop. The human who is not aware of dangerous driving situations needs some time and active effort to understand the critical situation and to determine how to safely maneuver the vehicle. This “bringing-back-in-the-loop” can be unsafe and cause risks of accidents or biased decisions. To get back into the loop requires the human to reestablish an understanding of the situation, either through establishing situation awareness themselves (a driver) or receiving explanatory information from the automation (a medical doctor receiving diagnosis suggestions). Out-of-the-loop problems are not new to human-system integration with complex technologies: nuclear power plants, modern aircraft, and many military applications exhibit high degrees of automation but still require humans for specific decision making. For such large systems, standard development processes have been developed to early on integrate the human role into the system design. Accidents happen if such processes are not followed such as the Chernobyl nuclear disaster in 1986 [5] and the Boeing 737 Max accidents in 2019 [6]. Also, users of end-user-devices have been observing these problems; for example, vehicle navigation systems are known to “occasionally” lure truck drivers to tiny mountain passes due to incorrect map data: it is difficult for even professional drivers to calibrate their trust for a system that seems to work right most of the time but then occasionally fails. As automation becomes pervasive, such experiences will multiply if not appropriately designed for calibrated trust.

The described challenges have been well recognized and are starting to be addressed worldwide. We start with an overview of available guidelines and approaches toward trustworthy AI in the next section. Then we identify remaining gaps that we address by introducing our approach in the subsequent section.

3 Current International Initiatives to address trustworthiness of AI

3.1 Guidelines and regulations

Governments and private companies address the challenges to develop ethical and trustworthy AI by proposing guidelines. A review shows that 84 guidelines on ethical AI were published worldwide until 2019 [7]. Analysing the authors of these guidelines reveals that private companies and governments seem to have a common interest in guiding ethical AI development. Private companies (22.6%) provided the highest number of guidelines, closely followed by governmental agencies (21.4%). Furthermore, most guidelines include similar aspects. The requirements transparency, justice & fairness, non-maleficence and responsibility are represented in a minimum of 71.4% of guidelines. The most frequent addressed requirement of transparency (in 86.9%) focusses on explainability, interpretability, data use and human-AI interaction. Figure 1 provides an overview of principles suggested by the AI guidelines.

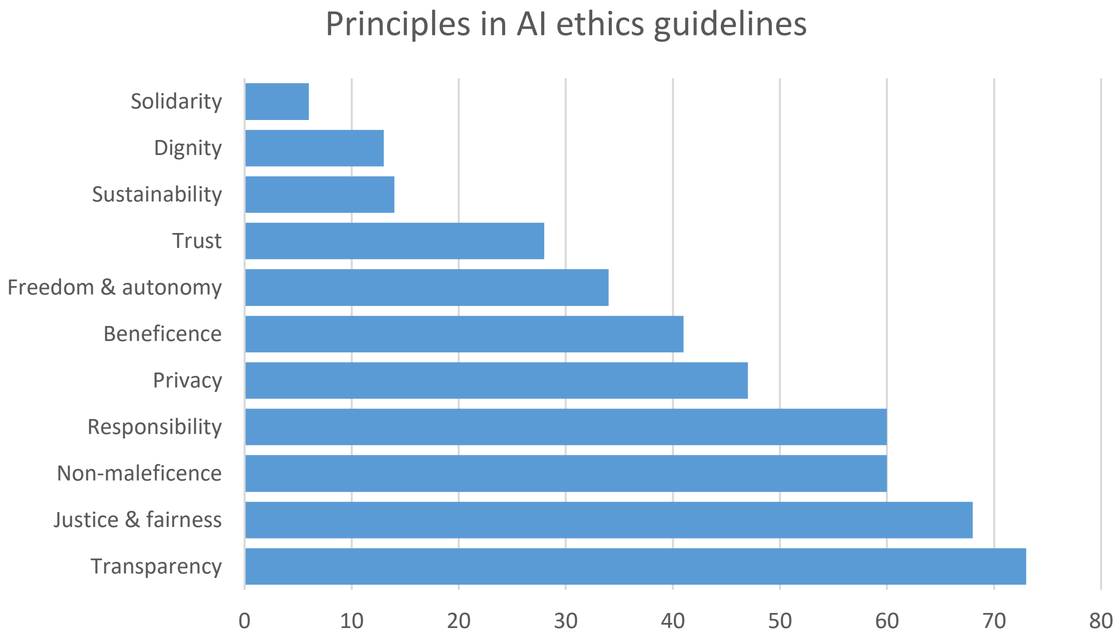


Figure 1 Numbers of most frequently addressed Ethical Principles

An AI ethics guideline that got high attention and serve as basis for further considerations of AI ethics (e.g., in [8], [9]) was provided by the European Commission´s High-Level Expert Group (HLEG) on Artificial Intelligence. The AI ethics guidelines focus on eight dimensions:

- **Human agency and oversight:** AI applications should support the user´s agency, autonomy, and decision-making.
- **Technical robustness and safety**: Technical robustness is central for precenting harm. AI applications should be developed to prevent risks and to ensure reliable functioning.
- **Privacy and data governance**: To prevent harm the privacy and date need to be protected and used data sets need to be of high quality.
- **Transparency**: AI application needs to be traceable and explainable. It should be transparent that an AI is in operation.
- **Diversity, non-discrimination, and fairness**: Diversity and inclusion needs to be ensured throughout the AI application´s life cycle.

- **Societal and environmental well-being**: To ensure fairness and prevent harm, the environment should be considered as stakeholder.
- **Accountability**: Auditability and reporting of negative impacts are important.

Building upon the EU ethics guidelines, the European Commission proposed regulations for high-risk AI. The proposed act includes ethics aspects and is currently in the European legislative process. [11]. The proposed legislation includes similar aspects of trustworthy and ethical AI and will make them mandatory for systems with high safety risks. Focus are applications in areas like biometric identification, critical infrastructure, or employment access. The perspective of mandatory requirements increases the urgency of AI industry to bring AI ethics in application.

The presented guidelines have a large area of applicability but lack implementation processes. For this purpose, several implementation processes have been proposed.

3.2 Implementation processes

One type of implementation process consists of checklists that allow AI developers a fast estimation of how well AI ethics are considered in their AI application. The AI ethics guidelines directly come with a checklist list for the ethics criteria (ALTAI; altai.insight-centre.org). The checklist allows developers to self-assess their AI application regarding ethics requirements. Based on the developers' responses, the assessment list provides explanations on what ethical aspects are missing. Similarly, checklists exist for software development applications like regression test selection [10].

While checklists are applied towards the end of development, the Eccola approach aims at starting ethical discussions throughout the development process. The approach summarizes different ethics guidelines to provide cards with easy understandable explanations and questions on ethical AI topics [9]. Eccola invites developers to discuss relevant topics throughout their development sprints. Concrete discussions outputs are documented, and the process aims at making developers aware of AI ethics.

The Z-Inspection approach takes the responsibility off the developers and involves interdisciplinary experts. The expert team follows a process to identify AI risks and to provide development with concrete recommendations. Within the Z-Inspection process, socio-technical scenarios are used to identify ethical issues and tensions. Furthermore, the results are mapped with the AI ethics guidelines.

3.3 Observed gaps in current initiatives

The reviewed guidelines toward developing ethical and trustworthy AI are very comprehensive and stand in no comparison to the number of processes to implement them. Even the available guideline implementation processes are limited in detail and depth. For example, while simple checklists are easy to apply by developers, they do not capture critical contextual information outside the developer's expertise. Also, the checklists are not embedded into a complete development that provides motivation of why a developer should care about the ethical outcomes of an AI system

Similarly, the ECCOLA approach raises important ethics related questions but does not per se prioritize the questions and does not provide success criteria. Also, other domain experts are not foreseen to participate which the limits the outcomes purely to the perspective of the developers.

Z-Inspection provides a process of how interdisciplinary experts develop recommendations for specific AI applications. However, specifications are required on how the recommendations are then actually implemented. We expect that considering the ethics recommendations that result from the

Z-Inspection in development bring a cost. In consequence, an organizational process to ensure the consideration of AI ethics in actual development is required.

The needs for an ethical AI development process that includes the organizational framework becomes apparent when analyzing the proposed EU AI act. While available methods for considering AI ethics are applied within traditional development, analysing the EU AI act shows the need for changed development processes. To provide an overview, we categorized a selection of regulations into the categories algorithm, development process and functionality (Figure 2) and complemented them by aspects out of the EU ethics guidelines that are not included in the AI act.

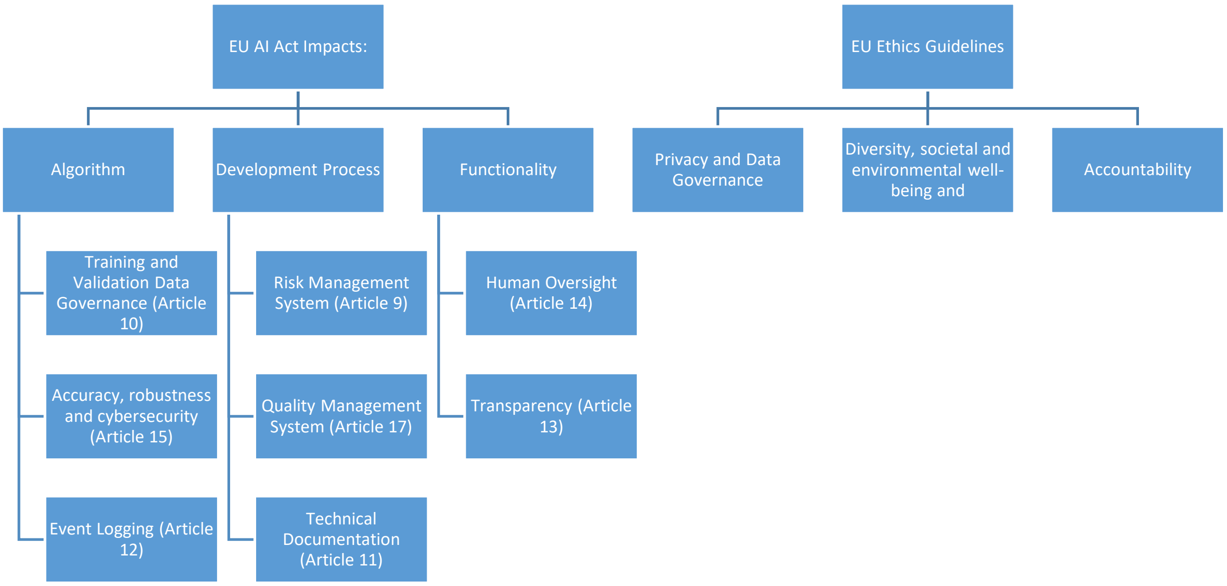

EU AI Act Impacts:
EU Ethics Guidelines
Algorithm
Development Process
Functionality
Privacy and Data Governance
Diversity, societal and environmental well-being and
Accountability
Training and Validation Data Governance (Article 10)
Accuracy, robustness and cybersecurity (Article 15)
Event Logging (Article 12)
Risk Management System (Article 9)
Quality Management System (Article 17)
Technical Documentation (Article 11)
Human Oversight (Article 14)
Transparency (Article 13)

Figure 2 Impact Areas of EU AI Act and EU Ethics Guidelines

The algorithm requirements bring the need for continuous engineering processes that do not end with fielding an AI system [12]. Algorithm requirements specify how AI models are developed, how AI is documented, how data is handled, and how the system´s activities shall be recorded. In traditional production, the manufacturers' active role ends with selling a product or, at the latest, when the product warranty ends. This was already extended by the latest software products that require security updates. The proposed AI act brings a further extension and makes risk and quality management throughout the system´s life-cycle mandatory for high-risk AI systems and thus has implications for R&D processes.

Functionality requirements bring the need to involve users in developing AI systems from early on. The upcoming regulations require high-risk AI to be designed transparently so that users can "interpret the system's output and use it appropriately". Furthermore, they shall be designed and developed so that they "can be effectively overseen by natural persons" (Human Oversight). In sum, functionality requirements enable the user to interpret the system´s output and oversee its activities. We argue that approaches of conducting user testing on prototypes and final systems are not sufficient per-se anymore because they often come too late to have a high impact on the product´s concept. Therefore, intensive involvement of stakeholders from the very beginning of AI development is necessary.

4 From Technology-Centered to Human-Centered Development of Smart Technologies

As outlined above, the developing trustworthy smart technologies requires shaping the development process toward the specific user and user context situations out of which ethical and trustworthiness issues only emerge. This can be difficult in traditional development environments where technology-centered development processes shape the role of users and their tasks, quasi as byproducts of the technology (A in Figure 3). This can result in “unbalanced” systems where the tasks or users may or may not be acceptable and trustworthy. Instead, what is needed for trustworthy development is shown to the right of Figure 3 (B) where the joined considerations of user, tasks, and technology, as well as task environment lead to a balanced system where trustworthiness and acceptable are part of the whole development process. The second process seems necessary to conform with the EU AI act.

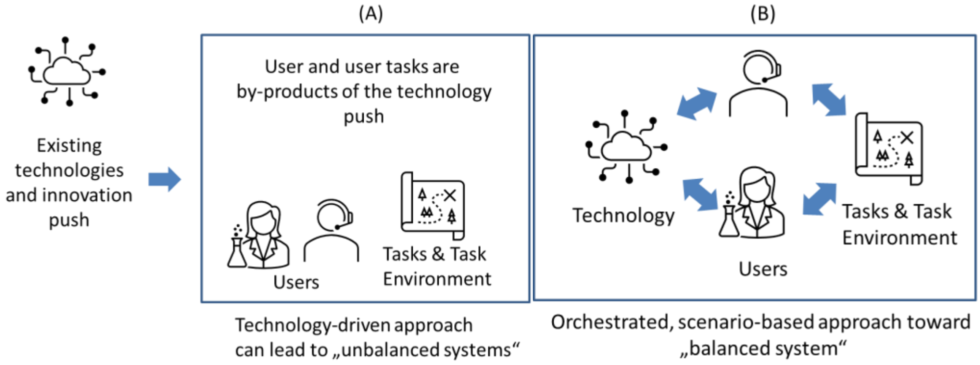

(A)
(B)
Existing technologies and innovation push
User and user tasks are by-products of the technology push
Users
Tasks & Task Environment
Technology
Users
Tasks & Task Environment
Technology-driven approach can lead to „unbalanced systems“
Orchestrated, scenario-based approach toward „balanced system“

Figure 3 Technology (A) versus Orchestrated (B) Process

To exemplify the differences between approach (A) and (B), imagine a developer developing a facial recognition system that recognizes criminals and handing it to the users, without knowing or clearly considering the possible consequences of algorithmic problems. Therefore, the technology is then transitioned to operational use (A). As result the users (and the larger public) find out that the system incorrectly categorizes minority members more likely as criminals than majority members. This is unintended and results in loss of trustworthiness on the societal level. In contrast, approach (B), the developing organization undergoes a detailed analysis of the use-situation and extract such risks of biases and addresses them as part of the technology development. This is, in a nutshell, what the EU AI act attempts to do.

Therefore, a key aspect to develop trustworthy system is to sufficiently analyze and address the specific use situation and involved users and tasks early on and allow this to shape the product design and development. Specifically, the analysis of the user and use situation should result in a set of trustworthiness risks, i.e. risks that if not addressed, may lead to loss of trustworthiness. These trustworthiness risks are managed in a life-long risk management process and thereby guide the product development and life-long operational process.

Moving from technology-centered to human-centered development methods requires an orchestration framework and process. These are described next.

4.1 Orchestrating the Development of Ethical and Trustworthy AI

To move from traditional technology-centered approaches to human-centered approaches, we came to realize that is necessary to not only first establish appropriate organizational structures but then also establish the necessary processes for them to interact and achieve trustworthy outcomes, see e.g. [13]. The situation of developing AI resembles the development of safety and security critical systems where many requirements originate from the use of the system in its real application context and are often not available at design time until explicitly brought into the process through specific analytic and research activities. And this is often new with AI system. The safety challenges of an airplane originate in the real world of flight operations and related reliability analyses and then shape the safety requirements of the components and the overall operation. Similarly, security threats in the real world bring the prioritizations needed to implement the appropriate security requirements. A similar situation concerns the ethical and trustworthiness requirements that are also only observable once the system has been fielded.

4.1.1 The Human Systems Integration Framework

Realizing trustworthiness means to “break down the silos of excellence" within which most normal technological developments currently occur. Traditional development organizations often focus their expertise on innovations at the technical level that knows little or nothing about the actual objective or mission: the experts of machine learning algorithm usually have no idea about the requirements of an end-user to understanding the outputs of the algorithm for his or her work environment. This is however critical for acceptance. Technological competences and knowledge are necessary but not sufficient prerequisites for ethical and trustworthy products.

To successfully bring trustworthiness and ethical requirements into the R&D processes early on, organizational structures and responsibilities need to be defined upfront. Therefore, the HSI framework postulates three interconnecting cornerstones: (A) an organization that is able to conceptualize and investigate the use of technology within a sociotechnical context, (B) a holistic development organization that is able to identify solutions in a strong multi-disciplinary effort, and (C) a life-cycle long learning and maintenance operations that addresses continuously changing aspects of the system. These cornerstones are linked via tools, processes, and standardized certification schemes that help to bound the solution space and aid collaboration and teamwork (see Figure 4).

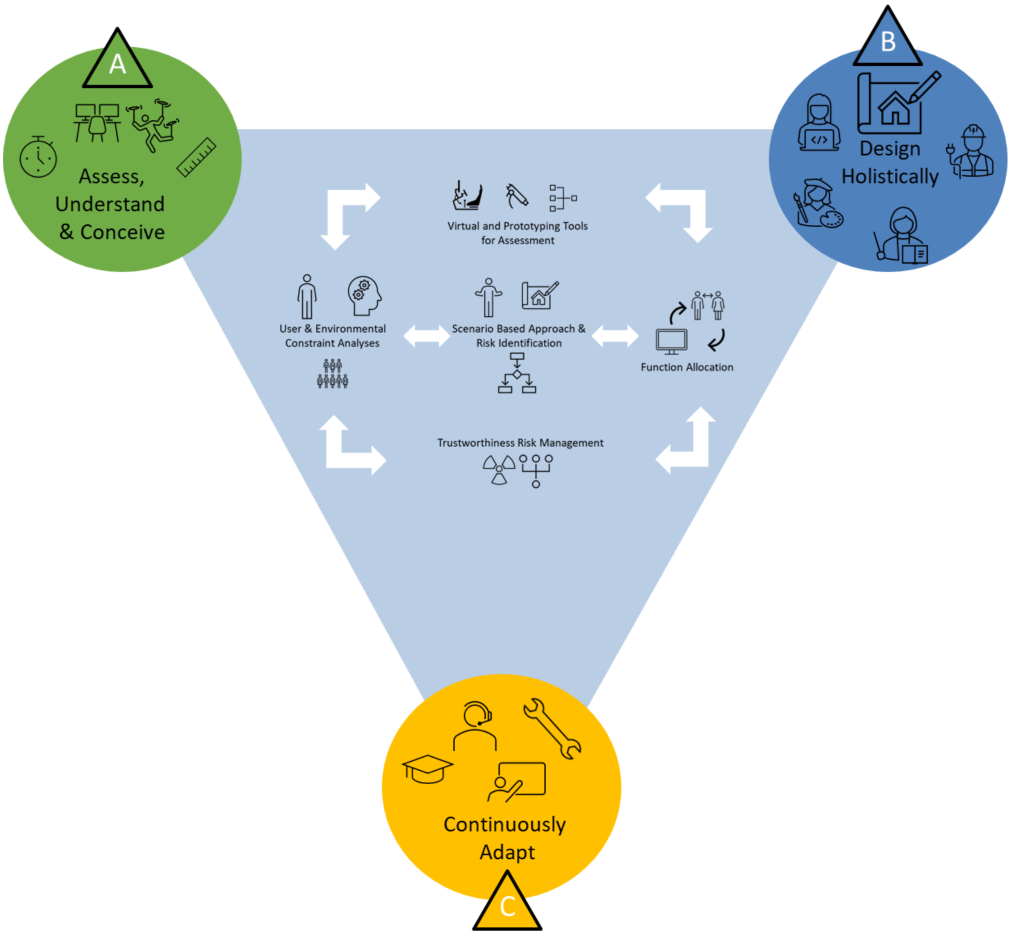

A
Assess, Understand & Conceive
B
Design Holistically
Virtual and Prototyping Tools for Assessment
User & Environmental Constraint Analyses
Scenario Based Approach & Risk Identification
Function Allocation
Trustworthiness Risk Management
Continuously Adapt
C

Figure 4 HSI Framework to Facilitate Trustworthy System Development Processes

HSI (Human Systems Integration) Cornerstone A "Assess, Understand, Conceive" is to provide the necessary information about the intended use situation for the development of smart systems that are to achieve sustainable acceptance and use. Such information includes the context of use, the goals to be achieved, as well as the user needs and important limitations of use and their situation and forms an essential starting point for the system design. Technical feasibility and cost-effectiveness are thereby concept-forming factors equal to the usage situation information; this is a novelty here. Such usage situation information is only available to the developers of today's systems to a limited extent. Usage situation information also includes characteristics of the user population and the tasks to be performed including criticality, responsibilities, and influences of the organizational environment as well as the work environment. In particular, organizational context and processes within which the system is used are important for design decisions, for example, to select appropriate methods of explaining smart technologies to the user. Data collections include observations, interviews, surveys, analyses, and especially virtual methods that allow users to make contextualized assessments (e.g., driving or flight simulators) as well as physical methods (e.g., Wizard-of-Oz studies). User and context are captured and translated into a high-level vision for how the system consisting of human, technology, and task & situation constraints could work.

HSI Cornerstone B "Design Holistically" translates the vision from cornerstone A into a holistic design of the system. The word holistic means here that orchestrated teams of multidisciplinary specialists work together to develop solutions across the various discipline-overarching dimensions. In so-called "living labs" products are co-designed to achieve a trustworthy, acceptable, and safe usage of the systems. This serves as a point of convergence across the disciplines and teams. Technical and cost factors act as limiting modulators. The challenge consists of making these larger contextual perspectives visible and overcome traditional isolated disciplinary hierarchies so that experts from different disciplines can effectively work toward such convergence. This requires sufficiently large, multidisciplinary research environments in a climate of positive holistic goal orientation and go beyond use and stakeholder abstractions as "matchstick men" (see (A) in Figure 3). Especially virtual simulation and modeling tools can support this process to combine the expertise of human factors, science and the various technical engineering disciplines.

HSI Cornerstone C “Continuously Adapt” consists of continuous adaptation and updating of products as well as the education of users during the life cycles of smart technologies are expected. System adaptations require detailed information about user and usage conditions. This requires a certain level of trust so that the user does not feel exploited or observed but sees himself as part of an improvement cycle. This also includes the possibility of user feedback which can not only promote user trust but also requires it. In addition to product adaptations, it is also important to promote the standardization of user knowledge and digital competencies in the form of standardized competence modules that enable users to find their way over time in what is otherwise perceived as a digitalization jungle. The creation of European training curricula for end-users, employees, and employers is a goal that must be initiated by technology developers, as this is where the critical information in the HSI process is available. The implementation of the digital competence modules in training curricula will then take place at the European level.

4.1.2 The HSI Process Model

Whereas the HSI framework postulates the organizational prerequisites for trustworthy and ethical smart technologies, how are these cornerstones stitched together into a working whole?

In Figure 5, the HSI process model shows the orchestration of the three cornerstones. Boxes 1 to 4 (green) indicate the processes in corner stone A “Assess, understand, and conceive”. This includes the risk management process. Boxes 5 to 9 indicate the activities of cornerstone B “design holistically”, and box 10 indicates the activities of cornerstone C “continuously adapt”. [14].

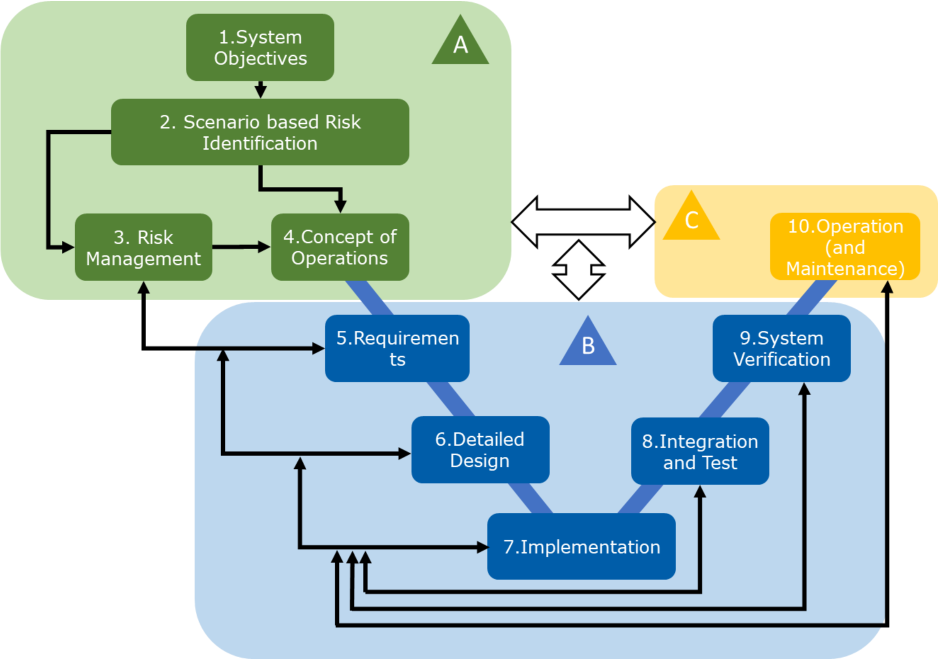


Figure 5 HSI Process Model for the Orchestration of Trustworthy systems

Critical in the HSI process model are the interactions between the three cornerstones to maintain the focus on the overall user experience concerning trustworthiness and ethical acceptability.

4.1.3 The Extraction of Trustworthiness Risks Using Scenario Based Methods

Scenario-based methods are commonly used in user-centered development efforts (e.g. [16]). Such methods can be used to identify risks by imagining the user within realistic and concrete environments risks [3]. As outlined above, the principles of ethical and trustworthy AI need to be contextualized in specific use conditions to become meaningful (see box 2 in Figure 5). Otherwise, they are too abstract to be able to derive specific requirements for implementation.

With a scenario we mean here a description of how an intended function can be accomplished under a realistic set of use conditions and stakeholder characteristics. A scenario thereby makes constraints visible that remain otherwise invisible. A risk consists of the description of a situation that, if it became real, would expose an undesired danger. We suggest that risks are at its core, formulated in simple sentences containing a precondition and a consequence, for example: “If a driver does not inform about his responsibilities, the risk for accidents is increased.”

The scenario description is the results of an analysis in which the ethic trustworthiness criteria are asked as questions: “What could that criteria mean in a specific condition for a specific user?”. In Table 1 we give examples of how this can be done around partial driving automation (SAE Level 2 [15]) that supports driving with longitudinal and lateral control, but leaves the driver fully responsible for monitoring the assistance and stepping to manual driving anytime when needed.

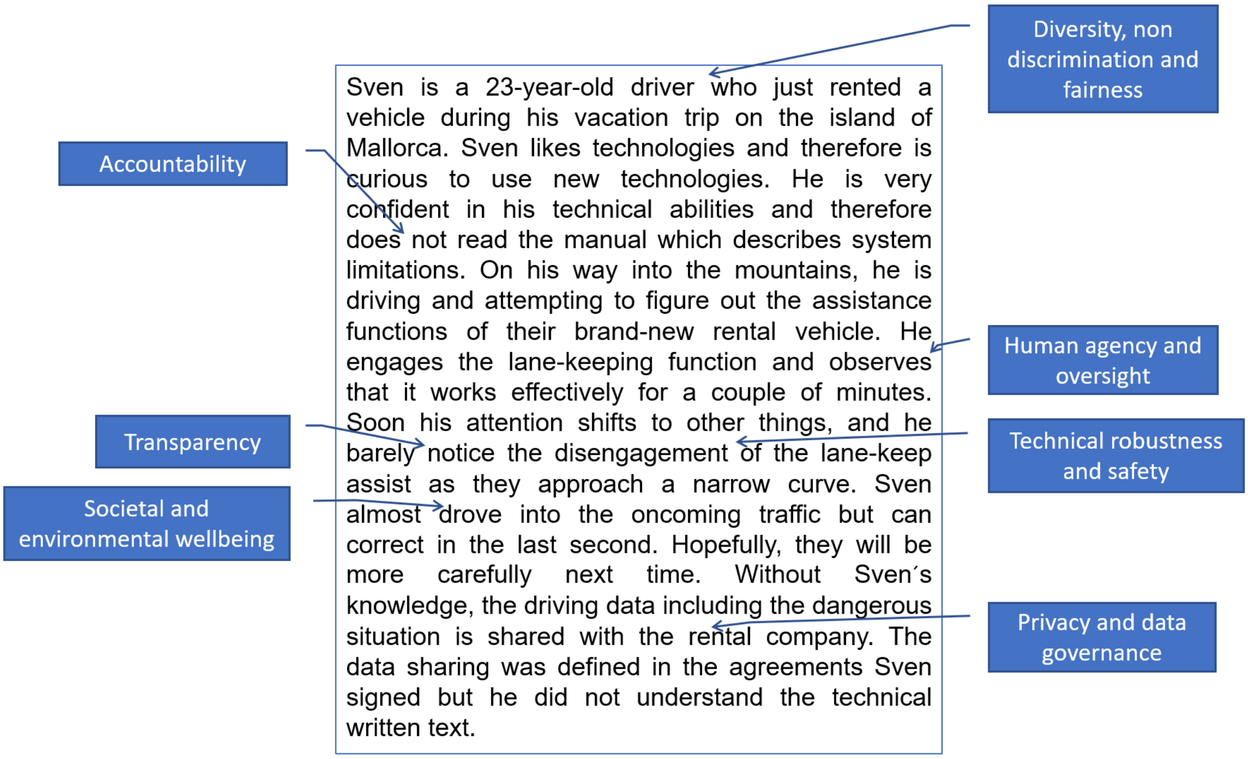
EU ethics criteria for trustworthy AI	**Contextualized EU ethics criteria**	**What does this criterium mean in a specific context?**
Accountability	The driver is solely responsible for the safety of the vehicle, the SAE Level 2 serves only as assistance. For this the driver has to clearly know the situations in that the partial driving automation is safe to use.	Real drivers sometimes may not be aware or willing to complete their assigned roles, resulting in foreseeable errors and violations. In reality, drivers often do not read the vehicle manuals where the responsibilities are defined. Therefore, the scenario should describe a realistic driver, not an ideal one, i.e., a driver who may not remember all of his/her responsibilities at all times.
Human agency and oversight	The human driver is responsible to monitor the driving environment and has to recognize when to take back control again.	Real drivers sometimes may not be able to complete their assigned roles of human agency and oversight, resulting in foreseeable errors. For example, a driver may get tired over time and not be able to keep up. Therefore, the scenario should reflect realistic driver behavior, not idealistic behavior (e.g. driver is sometimes

EU ethics criteria for trustworthy AI	Contextualized EU ethics criteria	What does this criterium mean in a specific context?
	Drivers should not use the system if it is not working reliably in a situation.	distracted, writes text messages on the phone, etc.).
Technical robustness and safety	The SAE Level 2 vehicle requires sensors to adequately detect the road environment, these sensors have to be kept clean, otherwise the partial driving automation will have problems to detect the street.	Technical robustness and safety may be accepted from a testing perspective but unacceptable from the operational perspective: e.g., when the user´s manual specifies that a system should only be used on dry roads, this shifts the problem from the system to the human who has to know the system limitations (which the user may not be aware of, willing to, or able to consider). The scenario should therefore consider technical robustness and safety from an operational user perspective.
Privacy and data governance	According to the EU AI act, data logs need to be kept for safety and quality assurance. Anonymized data on the partial driving automation are send to the manufacturer for improving functionality. As the data is anonymized it is not possible to relate them to a driver e.g., in case of an accident.	In reality, the logged data may be used for other purposes than initially intended, for example, to rate the driving behavior. The scenario should reflect how the collected data could be misused (a risk assessment how a human could be made aware about possible solutions).
Transparency	The vehicle´s automated driving modus is indicated to the driver. As drivers recognize the changed system state, they override or deactivate it to ensure a safe drive.	In reality, the driver is confronted with many vehicle status lights and indicators that may make it difficult for the driver to recognize the automated driving state indication. Therefore, the scenario should describe a holistic environment within a user interacts with the system (not just the to-be-designed system per se).
Diversity, non-discrimination, and fairness	An SAE Level 2 system should be able to adjust its headway to the lead vehicle based on user preferences	In reality, the adjustment of headway parameters can be hidden in many submenus and difficult to change. Therefore, the scenario should include different user groups (e.g., different age,

EU ethics criteria for trustworthy AI	Contextualized EU ethics criteria	What does this criterium mean in a specific context?
		experience, gender) to identify different preferences.
Societal and environmental well-being	SAE Level 2 should assist drivers and enhance their lives by increasing their safety and comfort. In consequence it intends to have positive impact on overall road traffic.	In reality, an ill-designed system may put the driver into stressful and uncomfortable situations. The scenario should explore these situations. (e.g., continuous enabling and disabling of SAE Level 2, unexpected or unacceptable driving behavior, etc.)

Table 1 Building Blocks mapping to Components and Scenarios

How can scenarios help to identify risks? Figure 6 shows a scenario that brings out EU ethics criteria in a concrete context and a concrete user. The scenario description focuses on a specific stakeholder. To ensure diversity, different scenarios representing different drivers (young/older, gender, …) are required in praxis.


Diversity, non discrimination and fairness
Accountability
Human agency and oversight
Transparency
Technical robustness and safety
Societal and environmental wellbeing
Privacy and data governance

Sven is a 23-year-old driver who just rented a vehicle during his vacation trip on the island of Mallorca. Sven likes technologies and therefore is curious to use new technologies. He is very confident in his technical abilities and therefore does not read the manual which describes system limitations. On his way into the mountains, he is driving and attempting to figure out the assistance functions of their brand-new rental vehicle. He engages the lane-keeping function and observes that it works effectively for a couple of minutes. Soon his attention shifts to other things, and he barely notice the disengagement of the lane-keep assist as they approach a narrow curve. Sven almost drove into the oncoming traffic but can correct in the last second. Hopefully, they will be more carefully next time. Without Sven´s knowledge, the driving data including the dangerous situation is shared with the rental company. The data sharing was defined in the agreements Sven signed but he did not understand the technical written text.

Figure 6 Example scenario to show the Link between Scenario and Trustworthiness Criteria

5 Conclusions

The key to creating trustworthy and ethical smart technologies consists of integrating humans and the system early on from the beginning of the system design process. Whereas current AI guidelines already prepare the principles and point to some methods that can be used by developing organizations to achieve this goal, these guidelines and principles also require an orchestrated development process and organizational structures and responsibilities that are currently not widely in place.

Multidisciplinary penetration of different fields of expertise must meet the establishment of responsibilities to leverage non-technical requirements to establish trustworthy and ethical use. Such requirements originate from an early definition and understanding of the use and context perspective out of which trustworthiness and ethical trustworthiness risks originate. These trustworthiness risks must be managed throughout the lifecycle of the product lifetime. Furthermore, scenario-based design methods seem promising to design systems for concrete users within specific use contexts. Orchestrating the development of use cases should stand on three organizational cornerstones with newly defined responsibilities and tasks. Trustworthiness risk management supports the life cycle of the product development. Such reorganization of the processes is seen as critical to break down the traditional silos of excellence of established engineering processes that focus on technological excellence and complexity decomposition.

Standard engineering processes in companies of today are often rather isolated in their specialties and separate from user and use contexts and simply often do not have the appropriate structures to bring the separate pieces to a harmonious whole. Similar to safety, trustworthiness is a holistic property of a system such that a small unforeseen event may render a system untrustworthy, for example when hacked or training biases disadvantage certain groups.

While it may be impossible to predict and address all possible trustworthiness issues, lots can be achieved when knowledge about the user and the use context are brought into the conception of a new smart system and risks are tracked throughout the design, development, and operation of the system. This is also anticipated by the EU rule on AI.

Therefore, in this white paper we have outlined the HSI framework that consists of design and organizational management processes develop trustworthy systems. Thereby the mentioned AI regulations do not have to be seen as burdensome checklists but, if appropriately used, can serve as motivators and enablers of successful innovations. We hope to have provided some initial insights in this white paper on how this can be done and look forward to your collaboration to further refine this white paper toward successful smart technologies.

6 REFERENCES

[1] J. D. Lee und K. A. See, „Trust in automation: Designing for appropriate reliance“, *Hum. Factors J. Hum. Factors Ergon. Soc.*, Bd. 46, Nr. 1, S. 50–80, 2004.

[2] High-Level Expert Group on Artificial Intelligence, „Ethics Guidelines for Trustworthy AI“, Brussels, 2019.

[3] D. J. Cahill, „CIHS White Paper: The Specification of a ‘Human Factors and Ethics’ Canvas for Socio-technical Systems“, S. 18, 2020.

[4] PwC, „Künstliche Intelligenz in Unternehmen: Eine Befragung von 500 Entscheidern deutscher Unternehmen zum Status quo - mit Bewertungen und Handlungsoptionen von PwC.“ 2019. [Online]. Verfügbar unter: https://www.pwc.de/de/digitale-transformation/kuenstliche-intelligenz/kuenstliche-intelligenz-in-unternehmen.html

[5] M. Salge und P. M. Milling, „Who is to blame, the operator or the designer? Two stages of human failure in the Chernobyl accident“, *Syst. Dyn. Rev.*, Bd. 22, Nr. 2, S. 89–112, 2006, doi: 10.1002/sdr.334.

[6] NTSB, „Assumptions Used in the Safety Assessment Process and the Effects of Multiple Alerts and Indications on Pilot Performance“, S. 13, 2019.

[7] A. Jobin, M. Ienca, und E. Vayena, „Artificial Intelligence: the global landscape of ethics guidelines“, *Nat. Mach. Intell.*, Bd. 1, S. 389–399.

[8] R. V. Zicari *u. a.*, „Z-Inspection \circledR : A Process to Assess Trustworthy AI“, *IEEE Trans. Technol. Soc.*, Bd. 2, Nr. 2, S. 83–97, 2021, doi: 10.1109/TTS.2021.3066209.

[9] V. Vakkuri, K.-K. Kemell, M. Jantunen, E. Halme, und P. Abrahamsson, „ECCOLA - a method for implementing ethically aligned AI systems“, *J. Syst. Softw.*, Bd. 20, Nr. 3, S. 111067, 2021, doi: 10.1016/j.jss.2021.111067.

[10] P. E. Strandberg, M. Frasheri, und E. P. Enoiu, „Ethical AI-Powered Regression Test Selection“, in *2021 IEEE International Conference on Artificial Intelligence Testing (AITest)*, Oxford, United Kingdom, Aug. 2021, S. 83–84. doi: 10.1109/AITEST52744.2021.00025.

[11] European Commission, „Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS“. 21. April 2021.

[12] R. Eitel-Porter, „Beyond the promise: implementing ethical AI“, *AI Ethics*, Bd. 1, Nr. 1, S. 73–80, Feb. 2021, doi: 10.1007/s43681-020-00011-6.

[13] G. A. Boy, *Orchestrating Human-Centered Design*. London: Springer, 2013.

[14] D. D. Walden, G. J. Roedler, K. Forsberg, R. D. Hamelin, T. M. Shortell, und International Council on Systems Engineering, Hrsg., *Systems engineering handbook: a guide for system life cycle processes and activities*, 4th edition. Hoboken, New Jersey: John Wiley & Sons Inc, 2015.

[15] SAE International, „Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles“, 2021.

[16] ISO, „Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems“, International Organization for Standardization, Geneva, Switzerland, ISO 9241-210:2010, 2010.

A. ABBREVIATIONS AND DEFINITIONS

Term	Definition
HSI	**Human Systems Integration**