AI and Ethics
https://doi.org/10.1007/s43681-025-00666-z

RESEARCH

Responsible AI, ethics, and the AI lifecycle: how to consider the human influence?

Miriam Elia[1] · Paula Ziethmann[2,3] · Julia Krumme[4] · Kerstin Schlögl-Flierl[2,3] · Bernhard Bauer[1,2]

Received: 15 November 2024 / Accepted: 22 January 2025

Abstract
Continuing the digital revolution, AI is capable to transform our world. Thanks to its novelty, we can define how we, as a society, envision this fascinating technology to integrate with existing processes. The EU AI Act follows a risk-based approach, and we argue that addressing the human influence, which poses risks along the AI lifecycle is crucial to ensure the desired quality of the model's transition from research to reality. Therefore, we propose a holistic approach that aims to continuously guide the involved stakeholders' mindset, namely developers and domain experts, among others towards Responsible AI (RAI) lifecycle management. Focusing on the development view with regard to regulation, our proposed four pillars comprise the well-known concepts of *Generalizability*, *Adaptability* and *Translationality*. In addition, we introduce *Transversality* (Welsch in Vernunft: Die Zeitgenössische Vernunftkritik Und Das Konzept der Transversalen Vernunft, Suhrkamp, Frankfurt am Main, 1995), aiming to capture the multifaceted concept of *bias*, and base the four pillars on *Education*, and *Research*. Overall, we aim to provide an application-oriented summary of RAI. Our goal is to distill RAI-related principles into a concise set of concepts that emphasize implementation quality. Concluding, we introduce the ethical foundation's transition to an applicable ethos for RAI projects as part of on-going research.

Keywords AI ethics · Responsible AI · Human mind · AI lifecycle · (Un-)fairness · Bias

1 Introduction

Artificial Intelligence (AI) will significantly reshape the world as we know it, and it is our responsibility to promote a beneficial outcome for everyone. Recently, the legislative landscape around this fascinating technology within the European Union (EU) was adopted, i.e. the AI Act. Generally, the novel EU regulation directs towards trustworthy AI (TAI), envisioning AI that respects fundamental rights, safety and health. The Act defines four levels of risk according to which intelligent systems are classified, and it is the provider's responsibility to implement and prove the required quality. Especially, high-risk domains such as medicine or education are highly regulated regarding the intelligent system's effects on the user, and its intended real world setting. The interface between implementation, and compliance assessment bundles together along the AI lifecycle, i.e. all processes, and design decisions that comprise the intelligent system's transition from conceptualization to reality [1]. Among other criteria, the identified stages are characterized by different humans, or stakeholders, depending on the system's intended purpose. As a result of mainly, but not

✉ Miriam Elia
miriam.elia@uni-a.de

Paula Ziethmann
paula.ziethmann@uni-a.de

Julia Krumme
julia.krumme@tha.de

Kerstin Schlögl-Flierl
kerstin.schloegl-flierl@uni-a.de

Bernhard Bauer
bernhard.bauer@uni-a.de

1 Faculty of Applied Computer Science, University of Augsburg, Augsburg, Germany

2 Center for Responsible AI Technologies, Augsburg, Munich, Germany

3 Moral Theology, University of Augsburg, Augsburg, Germany

4 Faculty of Liberal Arts and Sciences, Augsburg Technical University of Applied Sciences, Augsburg, Germany

Published online: 18 March 2025

limited to their decision-making, the human mind impacts the overall quality of intelligent systems, and their real-world performance. This human influence becomes evident, when analyzing TAI criteria, while one might argue that the multifaceted concept of *bias*, which we attempt to capture in the present paper, has the strongest ties to the human mind. It can result in undesired outcomes, and poses risks, that need to be anticipated and managed by the provider, and assessed by the regulator. Thanks to the omnipresence of *bias* and its inherent complexity, we argue that to date no suitable terminology to capture all relevant conceptual facets exists. Illustrating the transition of an abstract ethical concept to a concrete domain, we adapt *Transversality*, as introduced in [2], following [3] from media philosophy to AI and biased behavior. The concept can be traced back to the post-modern era, that is characterized by heterogeneity, and still influential today [4]. Globally, *Transversality* addresses the question how to shape a multi-dimensional society: "We live side by side in different worlds. Wolfgang Welsch postulates building bridges between them, which only a 'transversal reason' is able to do" [4, 1].

Concretely, we address the question how to promote a risk mitigating view on AI, shaped through *Embedded Ethics* [5] by design along the AI lifecycle, aiming for sustainability towards implementing responsible AI (RAI) [6], which is ethical (e.g. TAI), lawful (e.g. AI Act), and accountable (implemented in a manner that enables quality assessment). Complementing quality assessment, we focus on the human influence during system creation, resulting in the creation of a RAI mindset for contributing stakeholders, while we believe the outlined concept is extendable to more passive roles, such as the user. Within our proposed holistic approach, the four RAI pillars (*Generalizability*, *Adaptability*, *Translationality*, *Transversality*) are built on continuous *Education* and *Research*, which is crucial for a successful and long-term innovation integration. In addition, we refer to our proposed holistic approach as *ethical foundation*, which we aim to translate into an *applicable ethos* throughout the paper. Our proposed holistic approach is interpreted as a concrete realization of an *ethical foundation*, envisioned as potential candidate for RAI projects. Analogously, concrete materials to apply ethics during individual projects, as introduced in [7], for instance, comprise concrete realizations of the *applicable ethos*. All four interrelated pillars demonstrate desirable qualities for RAI, are profound in meaning, and complex to implement with respect to the multitude of existing AI use cases and application domains. In addition to considering bias, our proposed holistic approach promotes implementing systems that generalize well, while continuously adapting to new data and (unforeseeable) changes. We aim to promote a view on AI that automatically considers balancing stable and flexible behavior, which may result in trade-offs regarding concrete implementations of lifecycle design decisions. Therefore, reflecting upon *Generalizability* and *Adaptability* in detail is deemed valuable. As an additional perspective on the intelligent system's seamless real world integration, *Translationality* and *Transversality* each highlight different qualities. While the latter emphasizes social quality, including intricacies of related questions, the former sheds light on the concrete setting of the intended work environment, which may be characterized by particular hardware, existing infrastructures, and workflows. Together, we envision contributing the foundation of guiding directions for thought that enable a holistic approach to implementing RAI. Therefore, and addressing AI's inherent dynamics, the four pillars are grounded in constant *Education* and *Research*. Finally, our present contribution is embedded within our proposed RAI template on GitHub[1] as part of on-going research centered around our Methodology based on Quality Gates (MQG4AI) [8, 9]. The comprehensive lifecycle planning template, which currently is a work in progress, is envisioned to contribute to overall AI quality management (QM) focusing on RAI lifecycle design, information (& knowledge) management and linking, which we illustrate for TAI-related risk management (RM), a central component of QM, as starting point. Aiming to close the gap to applied ethics, we provide concrete materials for ethics training based on the DARE-method [7] for agile software development towards a RAI-mindset by design. Concretely, ethics are integrated once, generally, as basic RAI knowledge, and more specific in relation to the concrete application.

Section 2 introduces the context of RAI, how it relates to RM, TAI, and, focusing on *bias*, we highlight how the human influence plays a vital role when designing intelligent systems along the AI lifecycle. We transfer the concept of *Transversality* [2] to address AI and multi-faceted biased behavior in Sect. 3. Section 4 introduces our proposed holistic approach, its creation, and envisioned application in more detail. Section 5 discusses the challenge to move from ethical theory to practical implementation. Section 6 comprises the outlook.

2 Towards responsible AI (RAI)

Diaz-Rodriguez et al. introduce a holistic vision of Responsible AI (RAI), bringing together relevant concepts "[...] from ethical principles and AI ethics, to legislation and technical requirements" [6, 2]. Their three-fold approach envisions implementing AI that is ethical, lawful, and accountable. We accede with the authors, that "[...] in order

[1] https://github.com/miriamelia/MQG4AI/blob/main/README.md

to realize trustworthy AI that is compliant with the law, we advocate for the development of RAI systems, i.e., systems that not only make a responsible implementation that fulfills the requirements for trustworthy AI, but also comply with the AI regulation" [6, 19], aiming "[...] to attain [the] expected impact on the socioeconomic environment in which [the intelligent system] is applied" [6, 8]. The present section introduces the ethical, and legal context of RAI project planning, focusing on the the intersection of AI trustworthiness (TAI) criteria, the human mind, and risk management (RM), which comprises a key requirement of AI quality management systems (QMS) [10]. Finally, the impact of the human influence with respect to the multifaceted concept of bias, and (un-)biased behavior in relation to society, is highlighted, and we outline why the terms *bias* and *(un-)fairness* can be fragmentary.

2.1 The EU AI act, risk management and the AI lifecycle

The EU AI Act classifies four different levels of risk, depending on the intelligent system's intended use and scope regarding its impact on health, safety, and fundamental rights. We focus on high-risk systems, as concretized in Chapter III, Articles 6, 7 and 57, as well as Annex I of the AI Act [10]. Consequently, regulatory requirements apply, which are to be assessed in cooperation with an independent notified body, appending regulators as human stakeholders along the AI lifecycle. The implementation of compliant systems is carried out by the provider through a comprehensive AI QMS, as described in Article 17, that is also required to include a RMS, as defined in Article 9 [10]. The Act's ethical foundation "Ethics Guidelines for trustworthy AI" is introduced by the high-level expert group set up by the European Commission (HLEG) [11].

Definition of risk Overall, AI Act conform RMS are adapted to the AI lifecycle, comprehensive, adaptive, and continuous by design, include obligatory testing, and address post-market monitoring. *Risk* is defined as "the combination of the probability of an occurrence of harm and the severity of that harm" in Article 3 [10]. The concrete implementation of AI RM varies due to the specificity of each use case and the absence of standardized approaches for AI risk assessment (risk identification and evaluation), as well as the multitude of possible lifecycle design decisions (risk controls). Factors such as human influence, the use case itself (particularly data formats and tuning objectives), and the application domain create challenges for conformity assessment.

Addressing risk via trustworthy AI (TAI) The AI Act already provides a comprehensive list of possible AI risks, such as Article 14 on human oversight, for instance. From a practical perspective, the most generalizable level of AI-specific risk assessment can be addressed through the definition of TAI criteria. Their use case-adapted realization results in AI risk reduction. The NIST (National Institute of Standards and Technology) AI Risk Management Framework (AI RMF) introduces "[...] characteristics of trustworthy AI and offers guidance for addressing them" [12, 12]. They contribute a practical approach to realize overall AI risk management (RM) on a horizontal, and not domain-specific level, which comes closest to the RM interoperability framework, published by the Organization for Economic Co-operation and Development (OECD) [13], who contrast different approaches, among which the EU AI Act's Article 9, and ISO/IEC 23894:2023 *Information technology—Artificial intelligence—Guidance on risk management*.

Compared with the HLEG's TAI criteria, which are explained in a more practical manner through structured questions in the *Assessment List for Trustworthy AI (ALTAI)* [14], both perspectives are semantically very similar, but differently structured. It is worth highlighting that NIST visibly intersects the RAI dimensions through requiring *valid & reliable* outcomes, and considering *accountability & transparency* for the implementation of all other TAI criteria. Together, both perspectives on TAI form a solid basis for organizing important directions of thought. Figure 1 combines both views through the lens of the human influence from the implementation point-of-view. Safety, security and related concepts, such as resilience, which "[...] is the ability to return to normal function after an unexpected adverse event [...]" [12, 15], are tied to the human's state of mind, which becomes evident regarding phishing mails, or the evaluation of foreseeable states and possible harm during RM. Further, explainability and interpretability comprise approaches with the human recipient in mind, privacy and data governance lay the foundation for human autonomy, and fairness, as well as accessibility are closely tied to human user personas of the intelligent system. Finally, considering the broader context, based on e.g. the *Sustainable Development Goals*,[2] as defined by the United Nations (UN) for intelligent system development is crucial for long-term success. Overall, these directions of thought lead to a comprehensive RMS.

Finally, it is to be noted that TAI attributes are interrelated and characterized by tradeoffs depending on the respective real-world context. Their prioritization is domain and use case-dependent and includes the involved human along the lifecycle. For instance, the need for privacy and explainability may contradict each other.

Ethical, legal, and accountable dimension of RAI The *ethical* dimension of RAI is traced back to Recital 14a of

[2] https://sdgs.un.org/goals.

Combined View with particular emphasis on the Human Influence	HLEG High-level Expert Group, set up by the European Commission	NIST The National Institute of Standards & Technology (US)	
Human State of Mind, Reliability & Safety	Technical Robustness & Safety	Safety	Accountability and Transparency
Human State of Mind, Robustness & Security	Technical Robustness & Safety	Secure & Resilient	
Human Recipient, Explainability & Interpretability	Explicability (as part of their ethical foundation for TAI criteria) & Transparency	Explainable & Interpretable	
Human Autonomy, Privacy & Data Governance	Human Agency & Oversight, and Privacy & Data Governance	Privacy-Enhanced	
Human User, Fairness & Accessibility	Diversity, Non-Discrimination & Fairness	Fair – With Harmful Bias Managed	
Global Impact, Living Beings & the Planet	Societal & Environmental Well-being	Included in Appendix B & C as „global impact“	
Fidelity (HLEG: Technical Robustness) (NIST: Valid & Reliable)			

Fig. 1 TAI criteria fusing the structure proposed by NIST [12, 12], and the HLEG [11, 8], including a combined view that highlights the human influence from the development perspective. The notion *Fidelity*, introduced in [15] as fulfilling *completeness*, and *soundness* criteria in the context of Explainable AI was chosen to provide a human-centered view, and to summarize both perspectives under one term

the EU AI Act: "While the risk-based approach is the basis for a proportionate and effective set of binding rules, it is important to recall the 2019 Ethics Guidelines for Trustworthy AI", and emphasize that they "[...] should be translated, when possible, in the design and use of AI models", and incorporated in "the development of voluntary best practices and standards" [10]. The necessary AI lifecycle integration to realize these abstract concepts defining TAI on system level, including compliance assessment, which addresses the *legal* dimension of RAI, is carried out through a comprehensive AI QMS that is required to include a RMS (Article 9), as stated by the AI Act in Article 17 [10]. These systems are intended as an additional layer that can be appended to existing methodologies, if a shared definition of risk exists, while managing the quality of the system's real world performance, focusing on AI risk mitigation, and they are implemented through e.g. AI Act conform standards [16].[3] Finally, the *accountable* component of RAI intersects the quality of the lifecycle implementation with a "readiness" for compliance assessment, emphasizing practicality towards translationality. RAI systems promote "[...] auditability and accountability during [their] design, development and use, according to specifications and the applicable regulation of the domain of practice in which the AI system is to be used" [6, 18]. Therefore, it needs to be guaranteed that the processes, and design decisions that comprise individual AI lifecycles are reliable, and protect safety, fundamental rights, and health [10]. This necessitates a comprehensive AI literacy of contributing stakeholders, and, among others, continuous testing, and documentation strategies for implementation, as well as transparent compliance assessment.

In summary, the previously introduced perspective on TAI rudimentary links ethical principles that mirror fundamental rights with the need for accountable AI lifecycle implementations, while addressing compliance through providing guidelines for AI risk mitigation, and QM that can be systematized. Challenges arise due to AI's inherent complexity, as well as the required use case-specificity of concrete technical methods for RAI implementation. As of now, we are not aware of all RAI approaches for the multitude of possible application scenarios, and more research is necessary. The European Commission's standardization approach is introduced in [16]. Especially, negative biased behavior poses a global AI risk. It is characterized by a

[3] Refer to the EU AI Office for more specific information, especially this video: https://webcast.ec.europa.eu/risk-management-logic-of-the-ai-act-and-related-standards-2024-05-30?utm_source=substack&utm_medium=email.

multitude of different forms that are mostly re-traceable to human sources [17], as highlighted in the following section.

2.2 The special role of bias

Bias-identification and negative bias-effect mitigation are crucial for a trustworthy adoption of AI within society [18]. Challengingly, the multifaceted concept of bias is heavily influenced by different forms of human influence. Consequently, global measures to realize a transparent bias handling need to be developed and applied—under consideration of use case-specific differences. First, it is important to clarify what the multifaceted concept of bias means and how it relates with (un-)fairness. In its most basic essence and mathematical sense, bias could be translated simply with "a deviation." With respect to society, defining bias is not a trivial question, and multiple classification approaches exist. Generally, ISO/EC-TR 24027:2021(E) on Bias in AI Systems and Aided Decision Making [17] defines bias as the state of the system's "[...] input and the building blocks of AI systems in terms of their design, training and operation" [17, 3]. Consequently, bias as a state of (parts of) the AI system is different from resulting biased behavior within the system's real-world context, which does not necessarily lead to as negative perceived actions such as unfairness, for instance, but can have neutral or positive effects, as well [17, 5]. Further, intelligent systems are characterized by multiple sources of bias simultaneously [17, 7], and unwanted biases can cause the intelligent system to act other than expected if they stay hidden and untreated, which poses a risk [12, 12] [17, 5]. As a result, negative bias mitigation is centered around the definition of sources for bias along different components that need to be integrated with the AI lifecycle as part of risk management.

The following Fig. 2 compares three different perspectives on the meaning of bias, and, as will become apparent, the human influence is crucial. First, the structure introduced by ISO 24027 is presented in more detail, next, NIST's interpretation of bias (they published a comprehensive proposition for a "Standard for Identifying and Managing Bias in Artificial Intelligence" [19]) is highlighted, and finally, a more ethical perspective on bias by Fridman & Nissenbaum (FN) is illustrated. All of them highlight three main categories of bias, and we attempt to extract similarities, as well as analyze differences. Compared to ISO, NIST, and FN highlight "systemic bias" (or "pre-existing bias"),[4] while in ISO 24027 this perspective is part of the societal bias as a subcategory of human-cognitive bias [17, 9]. ISO's third category comprises bias introduced by engineering decisions, while NIST combines this technical perspective with bias in data sets under "computational & statistical bias", similarly to FN with their definition of "technical bias". Interestingly, FN promote "emergent bias" referring to the evolution of intelligent systems over time and the need for adaptability. Finally, all three definitions identify the human influence as root source for bias with slightly different realizations of structuring relevant sources. Also, NIST and ISO address how to handle bias from a more technical perspective, and introduce different mitigation strategies, while FN address "considerations for minimizing bias in computer system's design" [18].

2.3 Bias and *(un-)fairness*: why the terminology falls short

The concept of bias is multifaceted, its identification, as well as mitigation highly use case-dependent, and the bias-effect is evaluated in alignment with society's dynamics as wanted, or unwanted. To achieve an aligned bias-handling, it is of crucial importance to bring hidden biases to surface, which is related to the human mindset. Bias handling is preceded by bias source identification, and includes bias effect evaluation in accordance with society's dynamic perception, official legal regulations, and finally, an appropriate bias effect handling, which depends on the context. Consequently, neither *bias* nor *(un-)fairness* are suitable candidates to form a pillar of our holistic approach. The former is already very popular in usage, but often not interpreted or applied in a way that includes all relevant perspectives. The latter does not describe all possible undesired responses to bias, but focuses mainly on discriminative attributes from a societal viewpoint without explicitly considering negative technical bias effects, among others: "Fairness can be described as a treatment, a behavior or an outcome that respects established facts, beliefs and norms and is not determined by favouritism or unjust discrimination" [17, 3], a mechanism which in itself can result in unfairness. This definition highlights the concept's complexity. Evaluating and handling (un-)fairness depends on various contextual factors and its definition is fluid and evolves over time in alignment with societal shifts [17, 20]. Among others, considering temporal evolution is a necessary component when designing intelligent systems that, if continuously monitored and updated are long-term applicable. As a result of this complexity, and tailored to the human influence, instead of *bias* or *(un-)fairness*, we propose the translation of *transversal reason* [2] to AI and biased behavior, which results in *Transversality* as a fundamental pillar of our proposed ethical foundation.

[4] We believe this perspective is quite interesting, since possibly due to its complexity, omnipresence, and implicitness this source tends to be forgotten.

	Bias Category 1	Bias Category 2	Bias Category 3
ISO 24027	**Human-cognitive Bias** "Human beings can be biased in different ways" [17, 7]. They affect lifecycle design decisions at different stages. As a consequence, the system is biased by default, and human beings are a "root-source" of bias.	**Data Bias** "The data is a major source of bias" [17, 10]. It can be caused by human-cognitive bias, since data sets mirror the real world, and they way that they are structured is influenced by human beings.	**Bias introduced by engineering decisions** Model architecture, hyperparameters, manually designed features can be biased in different ways that are assessable on a technical level but executed by human beings collected real world data. "Data and human-cognitive bias can contribute to such bias" [17, 12], as a result.
NIST	**Human-cognitive Bias** The human as a source for bias reflects on lifecycle design decisions. This can be cause by an individual, as well as group dynamics. It is often implicit and can "[r]eflect systemic errors in human thought" [19, 9]. "These biases are omnipresent in the institutional, group, and individual decision-making processes across the AI lifecycle, and in the use of AI applications once deployed. There is a wide variety of human biases." [19, 9]	**Computational & Statistical Bias** Generally, a single sample is not representative of the ground population that comprises all relevant samples, which could mean the complete human population for some use cases. This scenario is not realistic. Consequently, "these biases are present in the datasets and algorithmic processes used in the development of AI applications and often arise when algorithms are trained on one type of data and cannot extrapolate beyond those data" [19, 9].	**Systemic Bias** NIST highlights the importance to reflect on risk posed by biases regarding lifecycle design decisions that are caused on organizational level [12, 40]. They "[...] result from procedures and practices of particular institutions that operate in ways which result in certain social groups being advantaged or favored and others being disadvantaged or devalued" [19, 6]. "These biases are present in the datasets used in AI, and the institutional norms, practices, and processes across the AI lifecycle and in broader culture and society" [19, 6]. Basically, this considers all biases that emerge from our current society and world in which human beings are socialized and the systems integrated.
Friedman & Nissenbaum	**Pre-existing Bias** Pre-existing bias refers to the biases that originate from social institutions, practices, and attitudes. These biases are embedded within the individuals who design and develop the technology, reflecting their inherent prejudices. As a result, these biases become integrated into the technology itself. [18, 333]	**Technical Bias** Friedman and Nissenbaum describe a technical bias that results from the technical constraints and design choices inherent in a technology. This type of bias can result from limitations in algorithms, data processing, or simplifications made during the design phase. As a result, these technical aspects may inadvertently favor certain groups over others. [18, 335]	**Emergent Bias** Emergent bias occurs when a technology interacts with its users and the context of its use over time. This type of bias is not present at the design stage but evolves as a result of changing contexts and new uses not anticipated by the designers. This evolution can lead to unforeseen and potentially biased results. [18, 336]

Fig. 2 Three different perspectives on bias (ISO [17], NIST [12, 19], and the foundational ethical perspective of Friedman and Nissenbaum [18], based on their work on *Value-Sensitive Design* that each highlight slightly different perspectives, while the human influence plays a crucial role throughout

3 How to address bias—introducing *transversality*

Mirroring the human influence, all "[d]ecisions that go into the design, development, deployment, evaluation, and use of AI systems reflect systemic and human cognitive biases" [12, 40]. As a consequence, we emphasize the need to integrate an ethical foundation transformed to an applicable ethos with RAI project planning and implementation aiming to guide stakeholders' mindset along the AI lifecycle. This section introduces *Transversality* as a comprehensive concept that, we believe is capable to open up *bias* in its entirety, contributing a fundamental pillar to our proposed holistic approach towards a RAI mindset in Sect. 4. We start with the concept's ethical origins, and its translation from application in media philosophy to the AI domain. For a practical interpretation of *Transversality* aiming to be integrated with RAI project management as foundation for continuous ethics training, refer to Sects. 4 and 5.

3.1 Conceptual background

In a postmodern and pluralistic society, the philosopher Wolfgang Welsch pursues a heterogeneous reason. His diagnosis of the times is that systems do not (or no longer) fit together, and the different forms of rationality do not fit together. Welsch appreciates the inherent logic and plurality of diverse rationalities and tries to think of a forum of

interconnections. This makes any claim to the truth of individual system logics more appropriate. *Transversal reason* plays an important role here. Despite all the diversity of forms of reason, he sees unity in the fact that transitions between these forms are possible.

> Reason must be more than such a formal general concept. In the past, it might have seemed obvious to regard such a concept - reason as a generic term and basic form of various rationalities—as sufficient, even desirable. [...] It is precisely through the multiplication and specialization of types of rationality that the task of reason has shifted, and its concept has changed: Reason is considered and today—in terms of plurality—precisely a faculty of connection and transition between forms of rationality. No longer cosmic, but earthly, no longer global, but connecting functions characterize the picture [2].

Reason is the transcending faculty in relation to understanding. All such reason that takes place in connections and transitions is referred to here as *transversal reason*. It is fundamentally distinguished from all principled, hierarchical or formal concepts of reason, all of which seek to comprehend or structure a whole and assimilate reason as understanding. *Transversal reason* articulates distinctions, links connections and drives disputes. It is to be understood horizontally and transversally; it does not overcome plurality, but only eliminates its aporias as a process form. The proximity of this reason to what is traditionally called *Power of Judgment*, in particular to its reflective form as described by Kant, cannot be overlooked. For transversal reason seeks common ground everywhere.

3.2 Bias, biased behavior and the concept of *Transversality*

Based on *Transversality*, we attempt to capture the essence of all relevant perspectives surrounding bias while aiming for a dynamic bias-awareness among stakeholders. In support of the inherent agility and dynamic setup of bias, biased behavior and fairness, we propose the transfer of the concept of *transversal reason* as defined by Welsch [2] to AI. While in the literature, some voices question Welsch's integrity regarding 'his novel invention' of *transversal reason*, for instance in [20], similarities with *rational reason* are outlined, we focus on the concept's inherent processes. Overall, we believe that the, in the following described dynamics surrounding *Transversality* combine, in addition to our multi-dimensional world, the fact that recent science focuses more and more on unstable structures (as AI or quantum mechanics) [4], which contributes to shaping our understanding of society as a whole, paving the way for reasonable actions.

3.2.1 AI and biased behavior

In its broadest sense, *Transversality* describes the process of evolving from theory to practice in a society that constantly adjusts theory, through addressing existing belief systems and measures towards their concrete realization or adjustment with respect to individual use cases. We are convinced that this profound perspective opens room for a continuous discussion as part of different formats for human interaction in AI project planning and implementation, which positively contributes to a RAI mindset. *Transversality*, originally published by Welsch [2], is translated to biased behavior in the context of AI based on the work by Sandbothe. She first adapted *Transversality* to media-philosophy, under the assumption that "[...] the World Wide Web proves itself to be a genuine medium of transversal reason" [3, 105], which is constituted by an accumulation of theoretically classifiable but disorderly arranged hyperlinks from a practical perspective when surfing online [3, 102]. Analogously, any AI model can be described as a medium of transversal reason based on its randomly constituted biased behavior that is all classifiable in theory [3, 105]. Biased behavior can harm society, if undetected and not addressed, which is why an on-going discussion is important. Sandbothe summarized *Transversality* in form of three related key points [3, 101], which we adapt to AI models and their biased behavior in the present section:

1. The constitution of rationality is characterized by an ineluctable disorderliness.
2. Reason is in principle capable of reconstructing and precisely describing this disorderliness.
3. It is only when reason productively analyses the subconscious entanglements of rationalities that it will be suitably equipped to solve contemporary problems.

The concept of *reason* has multiple stages of development, and the progression from *abstract* to *transversal reason* is desired for a functioning society that aims to solve contemporary problems. Further, the concept of *reason* to address *rationality* is executed by society, which, in itself is characterized by inherent dynamics. Finally, these assumptions are built on the axiom that *reason* in principle is capable to solve risks that can emerge with respect to unwanted biased behavior, in case of the present research. This conclusion is based on the fact that AI models, and their biased behavior can be characterized as a medium of *transversal reason*.

3.2.2 Concept translation—from theory to practice

In alignment with Sandbothe [3], we translate *Transversality* to the context of AI and biased behavior based on the following argumentation. First, the AI model is characterized as a *medium of transversal reason*. Next, the interplay of *reason* and *transversal reason* is introduced, embedded within society's dynamics. Finally, we conclude that this interaction describes the transgression of theory to applicable theory aiming to create desired effects on reality from a societal viewpoint—in a practical manner through the constantly shifting evolution of *reason* into *transversal reason*.

Transversal medium—*The constitution of rationality is characterized by an ineluctable disorderliness.* Different forms of biases constitute *types of rationalities* in form of the AI model's behavior and its (anticipated) impacts on the real world. The resulting biased behavior is characterized by a perceived randomness depending on the respective use case, and needs to be identified accordingly. Further, the existence of biased behavior is rooted in logical reasons that depend on various factors, have different sources, and appear at different stages along the AI lifecycle. As a consequence, and analogously to hyperlinks in the context of the World Wide Web [3, 102], biases are theoretically classifiable but disorderly arranged from an applied perspective, and we conclude that AI models can be described as a *medium of transversal reason*.

Reason—*Reason is in principle capable of reconstructing and precisely describing this disorderliness.* In theory, it is possible to develop and apply standardized methods that analyze the individual AI models to uncover and monitor their respective bias structure—in form of established and use case-specific methods that are applicable to groups of use cases, for instance. The individual realization of an intelligent system's lifecycle ideally includes the use case-adapted evaluation of and response to the model's identified biases and their effects fostering transparency, which enables RAI. This culminates in dynamic official requirements towards bias identification and handling that are first recorded in legislation and then need to be implemented on process level, i.e. the AI lifecycle and related information including stakeholders. Especially, since such systems are highly use case-dependent. The multitude of (non-)existing approaches aiming to realize and monitor the desired AI model behavior with respect to bias can be described as *reason* [3, 101], since we know that theoretically, the desired bias identification, and bias-effect handling are possible.

Transversal reason—*It is only when reason productively analyses the subconscious entanglements of rationalities that it will be suitably equipped to solve contemporary problems* Theoretical possibility itself does not equal objective bias-handling in the real-world. As of now, we are in the process of researching generalizable methods and standards regarding AI worldwide for the multitude of different AI use cases that exist, which includes bias identification and handling. This process is closely tied to (hidden) biases in today's world through the underlying data. Groups of AI use cases can be summarized in accordance with structural similarities, such as model type, data distribution and tuning objective, for instance. Commonalities that impact the appearance of bias need to be identified across application scenarios fostering a classification of existing methods, which simplifies their application. This shared research is a crucial step towards implemented RAI. In this regard, we are still relatively at the beginning stages of analyzing bias-entanglements with *reason* in a productive manner to continuously realize *transversal reason*. Further, based on our reasoning as society (which is biased, as well), an evaluation of the intelligent system's desired behavior occurs, which is constantly shifting depending on various factors. As a consequence, biased actions may become (un-)reasonable over time. Consequently, *Transversality* regarding biased intelligent behavior can be realized in form of RAI quality management strategies centered around the AI lifecycle. More precisely, bias-handling strategies should adjust well with AI's and society's dynamic behavior, which can be implemented by addressing the concrete realization and integration of continuous stakeholder- and context-adapted educational initiatives, for instance. Finally, with respect to the development of standardized and certifiable processes that foster RAI, intersecting education and research plays a significant role in bias identification, negative bias-effect mitigation, and handling.

Concluding, biases are unavoidable, their occurrence strongly use case-dependent and sometimes they are even desirable [17, 4]—but uncontrolled biased behavior is unreasonable and can result in severe consequences. It can appear in various forms based on different sources, which constitutes in a multitude of (groups of) use case-dependent, and preferably standardized methods towards implementing the current state of bias mitigation in society that overall adapt well to changes. These identified procedures constantly need to be (re-)aligned with official requirements and societal perceptions, as well as evolving AI model settings that are prone to variance. The described dynamics would benefit from a constant integration of discussing *Transversality* (and other pillars of the proposed holistic approach) as part of AI project planning to foster a general RAI mindset that fuels design decision-making, and process implementation. Consequently, we argue that on an even higher abstraction level, the described more dynamic than static interplay between *reason*, and *transversal reason* can be generalized across other concepts, use cases, and application scenarios.

4 Holistic mindset to design responsible AI (RAI)

As previously introduced, we emphasize the inclusion of an ethical foundation for RAI project planning, and as integral component of AI quality management systems (QMS). We envision an ethical foundation that, transformed to an applicable ethos, functions as a fundamental guide for contributing stakeholders along the AI lifecycle. Among other contents, the envisioned formation, which can be based on our proposed holistic approach, should include continuous bias-awareness trainings. They need to be adapted to the project-specific AI lifecycle, and its cycles of evolution over time in a practical manner contributing to continuous AI QM. Our, in the following introduced holistic approach, depicted in Fig. 3, is envisioned as an adaptable, and generalizable knowledge base to translate RAI for stakeholders. In addition to *Transversality*, the other pillars we identified are *Translationality*, *Adaptability*, and *Generalizability*, while their interpretation is built on the continuous interplay of *Research* and *Education*. At an abstract level, the pillars are exchangeable depending on the specific use case, while we believe integrating ethics with AI project planning, and *Education* and *Research* remain essential. First, our method to create the ethical foundation is introduced. We highlight considerations on the intended objective, which needs to be tailored to the respective target audience and application context. Then, we derive requirements to identify suitable concepts, aiming to define the ethical foundation's pillars, and its integration with AI project planning. The final step towards applicability is the preparation of materials, and we explore the ethical foundation's transformation to an applicable ethos in Sect. 5.

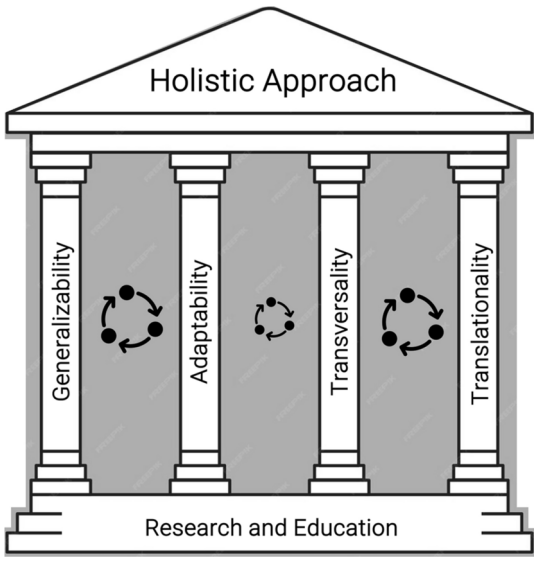

Fig. 3 Proposed holistic approach towards a Responsible AI (RAI) mindset for implementing intelligent systems that generalize and adapt well to new data, while performing a seamless real world integration—from a technical and social viewpoint. All pillars are interrelated and necessitate continuous consultation with respect to the AI's intended environment of use

4.1 Creation of an ethical foundation

Overall, the concepts that comprise our proposed holistic approach, that implements the ethical foundation, are envisioned to offer ground for continuous discussions, enable question-asking to foster a profound understanding, and at least broach all contents, and disciplines that comprise RAI to make them tangible. We are convinced that a thorough briefing of participating stakeholders as a constant part of AI lifecycle planning, and tailored to different stakeholder views is essential to realize RAI, and has, among others, a bias-mitigating and fairness-promoting effect on the entire system—until its decommissioning, and beyond. With respect to applicability, it is crucial to allow for an iterative consultation of ethical concepts in a consistent manner, and from different perspectives, integrated with project planning.

Objective and target audience Aiming for practicality during implementation, our proposed holistic approach comprises an adjusted representation of key aspects towards RAI focusing on involved stakeholders at the interface of development and regulation along the AI lifecycle for high-risk domains. In addition to developers, domain experts, and regulators, we highlight the role of RAI ethicists and researchers as part of the, for RAI lifecycles required diverse, and interdisciplinary teams. For instance, this could comprise the addition of a *RAI Person* who coordinates project-specific ethical knowledge management, among other duties at the intersection of a lawful, ethical, and accountable implementation.[5] In summary, our proposed ethical foundation is envisaged to be incorporated by all participating stakeholders along the AI lifecycle to align the foundation of their decision-making on implementation and organization level of RAI projects. Next, we derive suitable criteria towards applicability of selected ethical concepts based on objective and audience.

Concept definition–setup and requirements We identified the following key points shape the four pillars' creation to complete our proposed holistic approach towards RAI by design, while we believe the general structure is generalizable to other application scenarios.

[5] The target audience also includes addressing the user role, even though more passive by default (except for online learning scenarios, for instance), and it is anticipated during project planning for usability and ethical reasons such as safety.

- *Profound concepts for (continuing) discussion* The pillars and foundation of our proposed holistic approach should include relevant concepts that offer a comprehensive perspective on RAI, which is lawful, ethical, and accountable—tailored to the respective objective, audience, and use case. Further, identified concepts are intended to provide ground for on-going ethical discussions and open questions to understand different perspectives centered around the human influence, RAI, and society based on the continuous interplay of *Research* and *Education*. Depending on the context, our suggested four pillar-concepts are exchangeable.
- *Actionable approach* Our four identified concepts are envisioned to offer ground for promoting applicability of ethics in a domain- and stakeholder-adapted sense (development-view), so that they create awareness, which in turn is translated into action—through continuously training the human mind in a comprehensible way.
- *Short but rich in meaning* To avoid an overly complex setup, we propose four inter-relatable concepts, aiming to integrate continuity by design. Concretely, we address the question "how-to implement RAI" in form of four pillars, built on *Research* and *Education*, as depicted in Fig. 3. First, educating the intended audience on the four pillars' meaning with respect to the use case at hand is required to kick-start relevant ethical discussions envisioned to shape a RAI mindset as part of AI project planning in a continuous manner.
- *Iterative integration* Finally, the ethical foundation is required to be accompanied by a concept how to be integrated with the AI project planning, and QM processes based on identified scenarios. Among others, this comprises promoting AI literacy, and the anticipation of diverse user personas and interdisciplinary stakeholder roles to enable a comprehensive impact assessment. Rounding up the transition from the world of ideas to the world of action, we refer to our proposed RAI conceptualization template (MQG4AI) [8, 9] on GitHub (see Footnote 1), which includes an application-oriented ethics-section, as well as general knowledge on ethics in the context of RAI by design. It is introduced in more detail in the next section, and, among others, envisioned to contribute to organizing continuous *Research*, while providing relevant information to all stakeholders.

4.2 Proposed holistic approach towards responsible AI (RAI)

This section introduces our proposed ethical foundation towards RAI implementation, which is depicted in Fig. 3. As previously introduced, we attempt to translate the taxonomy presented in [6] towards lawful, ethical and accountable AI (RAI) for stakeholders that are actively contributing to the AI lifecycle, such as developers, domain experts, as well as auditors. Their influence shapes the concrete implementation of individual intelligent systems, as outlined throughout the paper. Therefore, we propose a holistic framework in form of an ethical foundation for RAI, that aligns with the AI Act's risk-based approach, emphasizing the ethical and societal dimensions of AI development. To translate RAI principles into actionable practices, we propose ethics training materials to foster interdisciplinary collaboration and a flexible, risk-aware development process along the AI lifecycle.

The framework is built on four foundational pillars: *Generalizability*, *Adaptability*, *Transversality* (see Sect. 3), and *Translationality*, providing a structured approach to guide stakeholders toward ethical and regulation-compliant AI lifecycles. Supported by continuous *Education* and *Research*, this approach promotes a "RAI mindset," embedding ethics into AI planning and development. Overall, we attempt to summarize RAI from an application-oriented viewpoint for contributing stakeholders that actively participate in/contribute to the AI lifecycle implementation. Therefore, we aim to summarize RAI-related values and principles in a concise summary of concepts regarding the quality of the implementation. All concepts are interrelated, and we envision a holistic mindset of contributing stakeholders that is continuously learning and evolving, while considering important questions related to transitioning AI into the real world. The four interconnected pillars are envisioned to lay the foundation, and therefore point out all relevant directions of thought surrounding RAI, highlighting ethical principles, the technical level, governance and policy-related decisions, social and environmental impact, as well as education towards awareness and AI literacy. For instance, the technical perspective on system accuracy and robustness is implied by *Generalizability*, and *Adaptability*, which equally broach concepts of security and interpretability. We are convinced, long-term these qualities are essential, and they are closely tied to e.g. (performance) evaluation metrics but include the monitoring-view early on. In addition, governance and social, as well as environmental impact are important considerations regarding *Translationality*, or when and how the system will enter the market. Finally, ethical values such as fairness, accountability, or transparency comprise the foundation of the entire proposed setup. We emphasize the inclusion of an ethical foundation with AI QM, closing the gap to regulation, and accountable AI. Additionally, *Transversality* emphasizes the role of bias and the question of fairness embedded within society's dynamics, and is envisioned to lay the foundation for further, holistic considerations when designing the AI lifecycle.

In summary, these four fundamental concepts are envisioned as basis for discussion to create awareness among contributing stakeholders towards a risk mitigating mindset for RAI design decision-making, lifecycle implementation, and project planning in general. Addressing AI's dynamic character, and use case-specificity, the successful internalization of the proposed holistic approach is based on continuous *Education* and *Research*. On-going *Education* and *Research* are relevant, since they comprise the flip-side of implementing AI regulation towards RAI, not only considering contributing stakeholders, but all humans that interact with intelligent systems. Intelligent systems need intelligent users—they are stochastic, opaque and evolving systems, and an interaction will always leave room for human interpretation to some degree for the multitude of possible use cases. In addition, AI's inherent fluidity necessitates continuous *Research* of suitable methods.

Finally, we focus on challenges in medical AI, which is, among others characterized by data imbalance and scarcity, complex domain knowledge, and high risks regarding the system's impact, while we believe that the proposed holistic approach is extendable across sectors and not only healthcare-specific.[6] In general, the domain of choice, and individual use case strongly impact AI lifecycle design decisions, and an understanding of the intended real-world setting, including a comprehensive analysis of related implications is an important enabler for RAI.

4.2.1 Generalizability

From an implementation view, *Generalizability* describes the capability of the intelligent system to generalize well to different modes of new instances it encounters [12, 14]. A model that performs consistently across different conditions, that promotes *Stability* and *trust*, is an important component towards RAI. However, implementing this is not a trivial question, and measures to continuously evaluate the model's generalizability need to be researched for the respective use case and application scenario. The data distribution during development for train, validation, and test sets needs to be made transparent, and related with the anticipated distribution of novel data after to-market release in the intended real-world setting. For accurate results, it is crucial to consider methodological pitfalls, such as a reliable application of performance evaluation metrics [22, 23] that enable trustworthy conclusions, the correct order of data pre-processing steps to avoid *data leakage* (the data splits are not independent of one another), and to consider a reasonable constitution of the data set's origins addressing *batch effect* (data samples are highly dispersed) [24], for instance. Methods such as cross-validation can help to evaluate *Generalizability*, if applied correctly [25]. As an additional *Generalizability*-layer, and with respect to facilitating RAI development, a standardizable application of identified, qualitative machine learning methods is crucial—especially, with respect to achieving long-term RAI. This includes establishing official bench-marking strategies, and test data, as intended by the AI Act regarding regulatory sandboxes in Article 57 [10]. The degree of reasonable *Generalizability* versus required use case-specificity offers ground for discussion, and can be realized through the identification of structural similarities. For instance, from a technical viewpoint, AI use cases and implementation approaches can be clustered based on criteria such as different data formats, or the type of model that is applied. Currently, much more research on the generalizable implementation of use case-adapted real-world RAI is needed to establish standardized procedures for (groups of) use cases, which is relevant for successful compliance assessment. As a result of AI's opacity, novelty and open questions for use case-specific implementations, *Generalizability* comprises a pillar of our proposed holistic approach to shape a RAI mindset for developers.

4.2.2 Adaptability

Closely tied to *Generalizability* but from a different angle (after to-market release) is "the adaptive behavior of a model as it is retrained on unseen data. This is an important model characteristic which should be considered in regulatory applications" [26, 1]. Aiming for *Flexibility*, data-driven, intelligent systems "[...] have the capability and need to adapt over time through continuous learning from [a dynamic] real-world experience after distribution" [26, 3]. Consequently, "AI systems may require more frequent maintenance and triggers for conducting corrective maintenance due to data, model, or concept drift" [12, 38], which results in novel challenges for RAI implementation and compliance assessment. Based on AI's inherent dynamics and opacity, the AI lifecycle is required to dynamically adapt to new technological advancements, as well as real-world context information, which is primarily mirrored by the data it encounters. Consequently, the utilized methods that comprise the project-specific AI lifecycle need to be designed in an adaptable, and flexible manner that fosters the ability to anticipate and react to changes from generalizable to use case-specific levels. As a result, processes to assess, and implement the required adaptability are crucial, and internal updates that can be triggered by events such as detected data drift, need to be considered from the start of project conceptualization for an efficient planning and realization of the RAI lifecycle. Also, more research is

[6] In [21], the authors propose to regulate foundation models, or general purpose AI similar to medical devices, referring to the unknown, and possibly high risks they can impose.

necessary, to better understand effects of related lifecycle design decisions. For instance, adding more data does not necessarily result in performance optimization, and "[...] different test sets possess different levels of challenge for prediction, demonstrating that the target test set appears to be the most important factor in performance" [26, 8]. With respect to project planning, all design decisions that comprise the intelligent system are envisaged to foster the individual project's implementation of *Adaptability*, which encompasses necessary dynamics for flexible future planning of the entire project with respect to its intended environment. We argue that discussing *Adaptability* as part of AI project planning and team meetings can promote a conscious communication of requirements for future-oriented RAI lifecycle management. Finally, *Adaptability* needs to be related with use case-specific trade-offs. For instance, regarding desired adjustments to the intelligent system, other aspects such as fairness or safety measures might need to be automatically reassessed depending on identified interdependencies. Overall, more research towards an adaptable integration of the intelligent system with an ever-shifting society is needed.

4.2.3 Transversality

Aiming to capture society's intricacies in their entirety as one term, we introduce *Transversality*, based on the multi-faceted concept of bias, to promote a responsible and adaptable negative bias effect-mitigating design decision-making mind set. Section 3 outlines the concept's transition from its ethical origins, and our proposed translation to RAI. As clarified in Sect. 2, systematizing the occurrence of bias, and understanding different sources is not trivial (disorderly arranged *rationalities*), but necessary. Consequently, it is our responsibility to participate in the process of shifting from *reason* to *transversal reason*—through the identification and implementation of negative bias-effect mitigating strategies that are continually aligned with the system's area of impact in the real world. *Transversality* provides ground to discuss society's dynamic bias-effect evaluation, shed light on the identification of (hidden) biases, and promote use case-adapted bias handling in relation to the human mind, the system's intended purpose, and underlying data. The overall aim is to develop and continuously update *transversal reason* in form of (standardizable) implementation processes that foster RAI, which include ethics training of the human mind through creating awareness. With respect to the concept's inherent tendencies, on an abstract level, *Transversality* can be interpreted as the process of dynamically (re-)defining applicable theory in alignment with a pluralistic and ever-shifting world. In addition to addressing existing belief systems, the concept is centered around the creation, and application of practical methods towards their concrete implementation in the real world. We believe this abstraction is extendable to other RAI-related ethical concepts: To promote a risk-aware, and -mitigating mind set, AI lifecycle design decision making would benefit from stakeholders that are trained to translate society's evaluation of AI trustworthiness into the intelligent system that they are creating—which is a logical consequence of the human mind, when profoundly, and consistently discussing ethical concepts, through creating awareness and asking questions.

4.2.4 Translationality

At the National Center for Advancing Translational Sciences at the US National Institutes of Health (NIH) *translation* is defined as "[...] the process by which a biomedical observation is turned into an intervention that improves health" [27]. Closing the gap to *Reality*, *Translationality* in the context of RAI emphasizes the process of transgressing from conceptualization and development to the intended real-world application deployment, maintenance, aiming for a seamless fusion with the existing workflow. To realize RAI, this transgression needs to be managed and understood depending on the individual use case and related domain knowledge. Consequently, we argue that understanding the long-term objective "real world integration" from the start is crucial to realize RAI. The developed intelligent system needs to be applicable and perform in the real-world, which often is not a trivial question. *Translationality* envisages to kick-start the discussion around practicality, real-world domain compatibility, and impact evaluation of the RAI project. It envisages to promote RAI's arrival in its intended workflow and not just in the realm of research. In contrast to *Transversality*, which emphasizes the question how society interprets RAI and its incorporation during design, *Translationality* focuses on the exploration of domain-specific implementation and compliance processes, as well as RAI knowledge, that are required to close the gap towards real-world applicability based on existing infrastructures. This aspect highlights all criteria of RAI, as discussed in Sect. 2 from an applied perspective starting from the intended domain and setting—"[...] in order to realize trustworthy AI that is compliant with the law, we advocate for the development of RAI systems, i.e., systems that not only ensure responsible implementation meeting the requirements for trustworthy AI but also adhere to AI regulation" [6, 19]. Among others, this includes domain-specific information on technology infrastructures, and scalability requirements, as well as ethical impact discussions of the intended use, including privacy or security requirements, and an on-boarding stage to educate users on relevant concepts, for instance. In addition to grounded methodologies

that implement quality management (QM). Depending on the respective use case, trade-offs need to be thematized, and domain knowledge translated between stakeholders as deemed reasonable towards sufficient interpretability of, and reliable interaction with the intelligent system.

4.2.5 Inter-pillar-relations

The four pillars are intended to be addressed continuously during AI project planning and execution, aiming to shape a shared mindset towards RAI, which enables use case-specific risk analysis, while offering enough openness to develop use case-specific realizations of desired criteria. They all comprise inter-related and fluid concepts regarding the intelligent system's transition to, and application in the real world. Their interconnectedness is illustrated in Fig. 3.

Generalizability, and *Adaptability* address more technical targets when designing AI, based on the technique's inherent opacity and dynamic character. AI's complexity necessitates prospective, and domain-embedded planning of risk controls with respect to all design decisions, AI lifecycle stages, and as a necessary contribution to conceptualization phases. Simultaneously, ensuring long-term, and adaptable RAI knowledge management, envisioning AI risk mitigation by design is considered a necessary step towards an optimized RAI implementation. *Generalizability*, and *Adaptability* imply requirements for applied existing and created methods, as well as lifecycle-adapted information management to assure the legally required quality. Further, both values close the gap to adapting implemented methods to the desired quality in the real world, since our reality is characterized by omnipresent diversity, in addition to moving fluidity, resulting in on-going changes and intricate evolutions. A RAI implementation needs to respect the movement and complexities of reality. Considering *Generalizability*, addresses this transition through the lens of the relation between data the model has seen, and novel samples it encounters, providing a stable, controllable constant—a continuous process that profoundly shapes the AI lifecycle's stability. Adding an additional layer to this foundation, *Adaptability* emphasizes the dynamics of the real world that the model will encounter in terms of evolving data. Implementing the desired flexibility supports a comprehensive understanding of the intended domain to enable prospective design decisions. Consequently, both concepts are closely related, with *Generalizability* comprising the foundation for positive *Adaptability*, while both necessitate continuous improvement surrounding the on-going interplay between stability and flexibility.

Moving to more concrete criteria of the intelligent system's transition into the real world, *Transversality*, and *Translationality* both shed light on different perspectives how reality, and society relate to the intelligent system. The latter focuses on the process of installing and integrating the AI system with its intended setting. This comprises questions regarding the concrete intended workflow, and how to blend user interaction seamlessly, impacting the design of explainable AI methods, for instance. Or, decisions on what packages or other resources to utilize for implementation, among other considerations that shape and are shaped by the intended environment of use. The former focuses more on the social climate of the intended real-world setting. *Transversality* references how society's dynamics impact the multi-faceted concept of bias directing towards the implementation of negative bias effect mitigation, while picturing the necessary dynamics to apply methods that realize the desired RAI lifecycle through the constantly shifting evolution of *reason* into *transversal reason* on an abstract level. Practicing awareness towards our personal biases is essential for any human being, and we are convinced that creating space to share these reflections in the context of AI project development (or, in general) will have beneficial effects on the quality of the lifecycle implementation. In summary, the reality, adapting to which is mirrored by *Translationality*, forms how considerations implied by *Transversality* will resemble, since they depend on the respective society, company, project team.

We propose these four pillars to form the foundation for a RAI mindset, and we aim for compatibility with existing AI project planning strategies for application-specific implementation scenarios. In summary, they are envisaged to function as a constant reminder and foundation for discussions starting during development how we, as a society aim to integrate RAI, what it means in general, and for the individual AI project. These powerful systems are usually implemented by a handful of human beings compared to the eight billion that exist—a huge responsibility, which needs to be addressed in a way that fosters trustworthiness and mitigates risks, long-term and in a continuous manner, including the human influence. Therefore, ethics training, interdisciplinary, and diverse teams are default requirements.

4.2.6 The role of research and education

For long-term internalization, the four pillars need to be continuously taught, discussed and updated from different angles addressing individual projects, so that ethical questions, and the ongoing discussion how to implement them are constant companions of the humans that contribute to the AI lifecycle within a particular domain. Consequently, Research and *Education* are closely related. They are included as fundamental drivers of our proposed holistic approach.

Thanks to AI's novelty and inherent dynamics, such as its opacity, stochasticity and use case-specificity, constantly updating AI knowledge in alignment with the broader research community, as well as the industry for RAI lifecycle implementations is crucial. Currently, a lot of use case-specific information on AI exists, scattered around the world in form of publications, best practices, and other formats. In Article 4, the AI Act states: "Providers and deployers of AI systems shall take measures to ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems [...]" [10]. In addition, education (which is closely linked to research) is interpreted as the flip side of regulation, or governance. Especially, in the context of AI, and the technology's inherent dynamics, for regulatory measures to become truly successful, the humans that interact with the intelligent system need to possess a certain degree of knowledge. For instance, regarding physicians and AI, there is a different level of profoundness depending on whether they adhere to a user or a domain expert role (or both) within an AI project. For instance, in [28, 29], the authors outline basic machine learning concepts tailored to application in the medical electrocardiogram (ECG) domain, which comprises a resourceful foundation for contributing domain experts. Vice versa, developers equally need to possess a basic understanding of the respective medical domain. Focusing on the medical user, less technical and more practical information on how to approach the intelligent system supports enhancing transparency and trust when interacting with the intelligent system, resulting in better patient care and respecting fundamental rights, safety and health [10]. This results in important considerations for on-boarding and monitoring of system performance within its intended clinical setting during deployment and maintenance. Knowledge of the internal functioning of the model is less crucial than providing a how-to-use approach, as physicians develop a mental model of how its influence should be evaluated during interaction with the intelligent system [30, 2]. For instance, the perception of physician-machine collaboration was attempted to be extracted and analyzed based on the integration of a real-time early warning system for patients with sepsis. The survey of physicians revealed, that especially the teaming perspective is decisive for successful adoption of the technology. The system was perceived as a competent "second pair of eyes" [30, 2], which, among other things, assists in organizational activities, such as prioritizing patient visits. Generally, reasonable levels of knowledge depending on the human's role need to be identified, and knowledge transformations executed. With respect to public large language models (LLM) such as ChatGPT, for instance, this basic knowledge may include concepts such as *hallucinations*, *scheming* and even a basic understanding how DNNs work, which ideally is taught in school.

Finally, in light of on-going global advancements in AI regulation, means to organize, structure and assess individual RAI lifecycle realizations need to be and are being developed. Methodologies that offer insights how to plan, and organize AI lifecycle processes can be found in [31] across sectors, and in [32] for medicine, for instance. As part of ongoing research, our next publication is submitted and about to be published as a preprint, we introduce our Methodology based on Quality Gates (MQG4AI), which is initially introduced in [8], and partly illustrated in [9] for a fictional use case addressing reliable performance evaluation metrics selection for multi-label classification in emergency medicine. Following principles from Design Science Research [33], i.e. the continuous communication of an abstract design knowledge base and concrete use cases, we focus on AI system-specific lifecycle information (& knowledge) management for project-specific application scenarios and decentralized RAI design knowledge organization alike. The generalizable template for high-risk AI, and customizable Quality Gate-format relate implemented concepts with regulatory requirements, focusing on AI trustworthiness-based risk management and AI system-specific information linking towards reliable lifecycle processes and design decision-making. The envisioned RAI-template on GitHub (see Footnote 1), which is a work in progress, incorporates our proposed holistic approach, and we envision to provide application-centered ethics-training, and -education as part of RAI project planning to all stakeholders by design. This includes the addition of materials for self-organized ethics-sessions, as discussed in the next section.

5 Discussion

Our main focus is the transition from research to reality for RAI. As a consequence, in the present paper, we aim to provide a practical approach for a simple and independent integration of ethical topics during project planning to shape a RAI mindset for contributing stakeholders. This adds to the previous section, where we suggest the integration of an ethical foundation, realized with our proposed holistic approach for RAI, with AI project planning, which is exemplified through MQG4AI. Further, at the Center for Responsible AI Technologies, we work with an innovative approach, "Embedded Ethics and Social Sciences" [5], to integrate ethics throughout the development process. This approach is embedded in an extensive research program that we will not describe in detail here. For more information see f.e. [34]. In the present section, we outline subsequent ethical topics, and discuss concrete materials, as well

as a structure for independent ethics training based on the DARE-method [7]. Essential materials can be found here.[7]

For instance, in agile development, the *product owner* acts as the primary interface between the product team and various stakeholders, including consulting, passive, external, or other active participants regarding the project. A key responsibility of the *product owner* is to effectively communicate the respective vision to ensure alignment and shared understanding across all involved parties. We propose to relate this role with knowledge on RAI. Figure 4 illustrates the agile development process, and at what stages ethical considerations within the team become particularly relevant. (A) focuses on the "what" of development, addressing the goals, requirements, and overall direction of the product. (B) emphasizes the "how," concentrating on the processes, methods, and approaches used to achieve the goals. Finally, (C) represents a moment of reflection, conducted before the start of a new iteration, to evaluate and adapt based on ethical and practical insights gained. This structured approach ensures that ethical considerations are woven into the agile workflow, promoting thoughtful and responsible development practices.

In the context of RAI, "[...] the difficulty in moving from principles to practice presents a significant challenge to the implementation of ethical guidelines" [35, 1]. Current frameworks tend to struggle with practical applicability, which is a necessary feature for our fast-paced economy. As introduced, we illustrate a possible application of our proposed ethos based on the DARE-method, which is currently being tested for agile software development, and can be integrated with familiar procedures [7]. To illustrate the desired practicality, we adjust a simplified version of the DARE-set to our setting. Among others, this includes guiding questions, flash cards, and a user manual, aiming to provide a solid foundation for independent, and decentralized ethics training in the context of RAI development. DARE, as depicted in Fig. 5, highlights critical topics such as training data quality and setup, the role of the human-in-the-loop, and explainability as essential components of RAI. These topics serve as anchors for understanding and addressing ethical, technical, and practical challenges in AI development. By focusing on these foundational (abstract) concepts, a comprehensive framework for RAI discussions emerges, constantly evolving in alignment withe project's cycles. When combined with specific flashcards, this framework creates a robust basis for transferring the principles and applicability of RAI to individual projects. Overall, this structure fosters meaningful conversations and promotes the integration of RAI values into diverse development contexts.

Other interesting scenarios surrounding RAI ethics sessions within a specific context could include Lego Serious Play as foundation for ethical discussions. A draw-back would be the lack of independence, since a skilled session

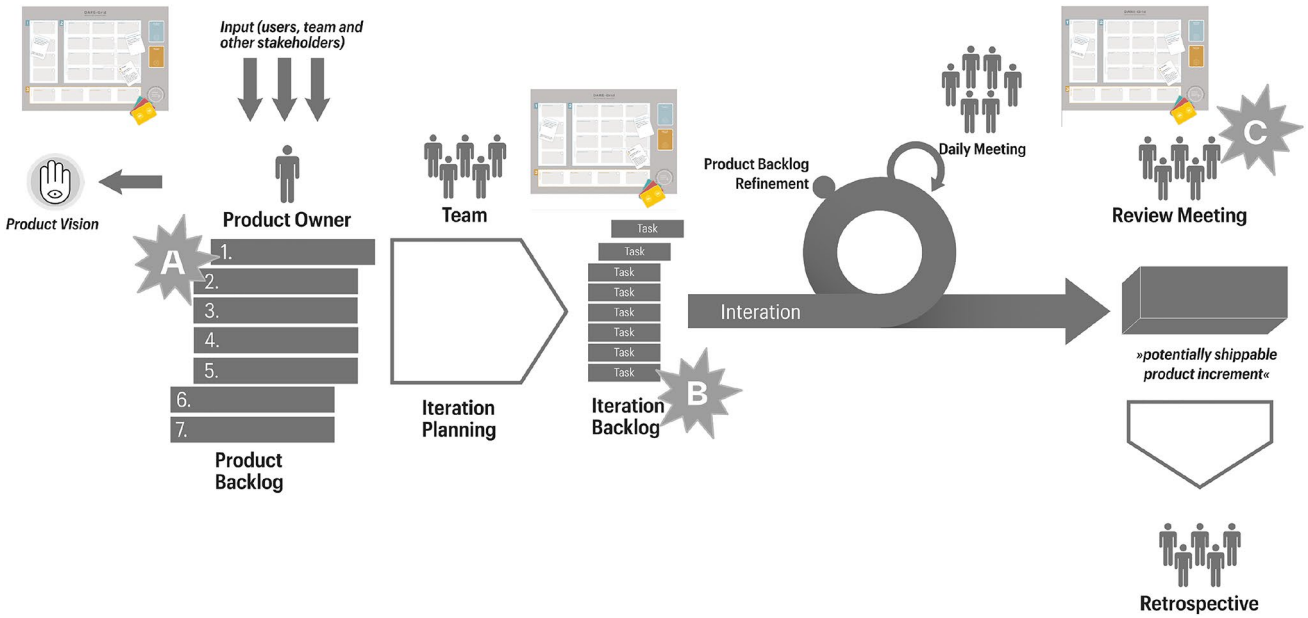

Fig. 4 Integration of DARE [7] with the agile process, towards *embedded ethics* [5]. The DARE method, which involves working with a grid and cards, can be meaningfully integrated at specific points of (agile) development. These points (A, B, and C) are where ethical considerations become particularly relevant. In addition, agile development highlights the iterative and evolving character of project-specific (RAI) knowledge. The DARE-grid is depicted in more detail in Fig. 5

[7] https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/1_System/Application/Ethics_Specific/Ethics_Specific.md

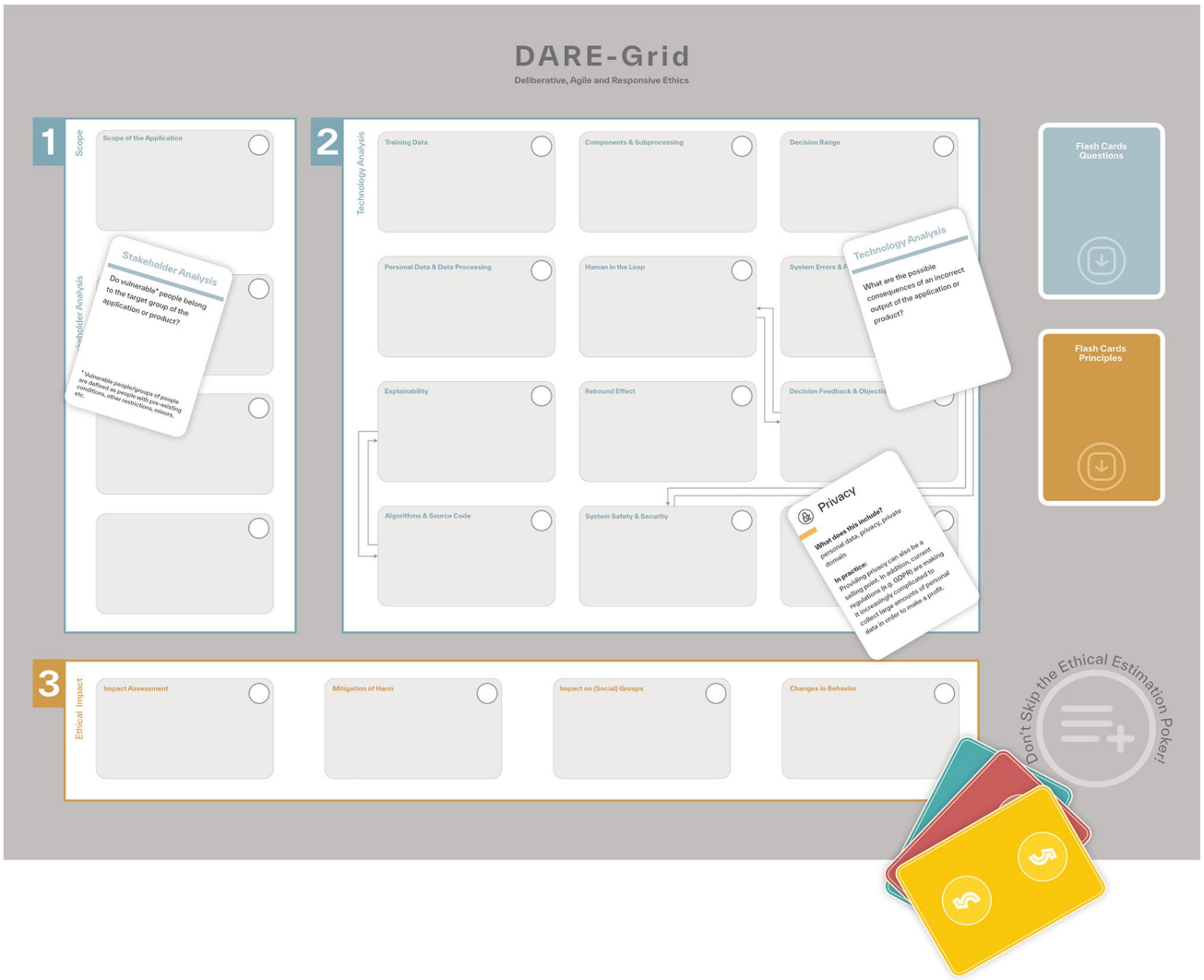

Fig. 5 Incorporating elements from *gamification*, the DARE-method [7] provides a solid approach to apply ethics during (agile) development projects

manager is needed for a fruitful experience. To still benefit from the positive effects of facilitating communication through self-made Lego models, such sessions could be included once or twice a year with an external professional, for instance, which simultaneously can have enhancing effects on the team spirit. Our vision is the provision of materials in such a way, that discussing the ethical foundation can become an integral part of RAI projects. In addition, we would like the suggested approach to be an agreeable experience for participating stakeholders that results in interesting AI ethics consultations, and we follow the 'gamification approach', aiming to create a fruitful learning environment. Concretely, we envision to enable the internalization of the individual pillar's meaning, and their interrelations, as well as to fuel open discussions on topics related to AI ethics or other, as reasonable identified concepts surrounding individual projects, while fostering a learning environment towards RAI literacy.

Additional consultations on brainstorming ethical concepts related to RAI, which support our proposed holistic approach, can be found in the literature. Complementing the introduction of RAI in [6], we recommend [35], where the authors extract RAI ethical recommendations including a medical use case example. Further, we would like to highlight ethical discussions surrounding (semi-)automation of AI development, application, and compliance assessment. We refer to [36], where the author discusses the question when a decision is automated and how to approach different workflows in a human-centered manner related to current regulation in Europe, i.e. the AI Act, and General Data Protection Regulation (GDPR). In general, we believe that optimizing towards a responsible semi-automation that is built on responsible humans-in-the-loop will benefit AI's arrival

in society—if emerging risks are identified and addressed properly, starting with creating awareness among contributing stakeholders along the AI lifecycle towards RAI.

6 Outlook

In the present paper, we propose the creation of an ethical foundation, and its transition to an applicable ethos, envisioned to be integrated with RAI project planning. We aim to enable the implementation of RAI, focusing on the human influence of contributing stakeholders along the AI lifecycle, which poses risks if not addressed. We propose the four pillars *Transversality*, *Translationality*, *Generalizability*, and *Adaptability* of our holistic approach, which are based on *Research*, and *Education* towards continuous, decentralized RAI knowledge management. Introducing *Transversality* in relation to the multifaceted concept of *bias*, we attempt to illustrate the transition of an abstract ethical concept to its interpretation in the context of RAI. Bridging the gap to applicability, we provide a possible approach that is intended as a foundation to shape a RAI mindset[8] as part of RAI lifecycle planning towards quality management by design. Further, we believe the proposed method in Sect. 4, which explains how we create the holistic approach, is extendable to other scenarios. We also believe the provided materials, based on, for example, DARE [7], can be tested in and adapted to various contexts, including (higher) education for the next generation of contributing stakeholders. Looking ahead, it is possible to expand the approach to test DARE, and to measure the long-term effects of an applied ethos during RAI projects, by complementing case studies with the development of a multi-level scoring system. This score would aim to measure quality from multiple perspectives, encompassing technical, ethical, and societal dimensions. Such a system could provide a comprehensive evaluation framework for assessing AI systems and their alignment with RAI principles. Additionally, this scoring framework could be used to compare the long-term impact of ethics training against a control group that has not implemented methods like DARE. This evaluation would not only assess the effectiveness of DARE, but also highlight the role of ethics in fostering awareness, accountability, and ethical decision-making throughout the AI lifecycle. Such insights would further validate the importance of structured ethical training and provide actionable feedback for iterative improvement.

Finally, we are convinced that promoting a global RAI mindset, which creates awareness for the responsibility of actions, and aims for life-long learning will have beneficial effects on many areas, thanks to the educated and aware human-in-the-loop. We hope to design our contribution in a way that offers substantial input for RAI ethics, and ethical embedding throughout the AI lifecycle, while also inspiring further ideas within concrete application scenarios. As part of on-going research, we equally emphasize assembling more empirical knowledge in the real world to deduct RAI design knowledge. For instance, in form of ethical training for different target groups, and overall as integral part of RAI software development, fostering the exchange of RAI development ideas. We all share one planet, and we believe that promoting awareness of the intricacy and responsibility of our actions, that shape the world of tomorrow is crucial—especially, regarding new technologies.

[8] https://github.com/miriamelia/MQG4AI/blob/main/MQG4DesignKnowledge/1_System/Ethics_General/Ethics_General.md

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Miriam Elia, Paula Ziehtmann, Julia Krumme, Kerstin Schlögl-Flierl, and Bernhard Bauer. The first draft of the manuscript was written by Miriam Elia and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL.
Conceptualization and writing of this work was partially funded by the German Federal Ministry of Education and Research (BMBF) under reference number 031L9196B.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no Conflict of interest.

References

1. ISO: Information Technology—Artifical Intelligence—AI system lifecycle processes—ISO/IEC FDIS 5338:2023(E). Standard, International Organization for Standardization, Geneva, CH (2023)

2. Welsch, W.: Vernunft: Die Zeitgenössische Vernunftkritik Und Das Konzept der Transversalen Vernunft. Suhrkamp, Frankfurt am Main (1995)
3. Sandbothe, M.: Interactivity - hypertextuality - transversality a media-philosophical analysis of the internet translated by a. inkpin. Hermes. J. Linguist. **24**, 81 (1999)
4. Widl, M.: Jean-françois lyotard der widerstreit. ZPTh Zeitschrift für Pastoraltheologie **43** (2023)
5. McLennan, S., et al.: An embedded ethics approach for AI development. Nat. Mach. Intell. **2**, 1–3 (2020). https://doi.org/10.1038/s42256-020-0214-1
6. Díaz-Rodríguez, N., et al.: Connecting the dots in trustworthy artificial intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regulation (2023) arXiv:2305.02231 [cs.CY]
7. Krumme, J., et al.: Never mind the codes of conduct. Dare you to tackle ethics in software development for ehealth. Stud Health Technol Inf **316**, 2–6 (2024). https://doi.org/10.3233/SHTI240330
8. Elia, M., Bauer, B.: A methodology based on quality gates for certifiable ai inmedicine: towards a reliable application of metrics in machine learning. ICSOFT (2023). https://doi.org/10.5220/0012121300003538
9. Elia, M., et al.: Towards certifiable AI in medicine: illustrated for multi-label ECG classification performance metrics. In: 2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS), pp. 1–8 (2024). https://doi.org/10.1109/EAIS58494.2024.10570023
10. FLI: Eu AI act explorer. Online, Future of Life Institute (FLI) (June 2024). Accessed 10 Oct 2024. https://artificialintelligenceact.eu/ai-act-explorer/
11. Commission, E., Communications Networks, C., Technology: Ethics Guidelines for Trustworthy AI. Publications Office, Belgium, Brussels (2019). https://doi.org/10.2759/346720. https://data.europa.eu/doi/10.2759/346720
12. Tabassi, E.: Artificial intelligence risk management framework (ai rmf 1.0). Framework, National Institute of Standards and Technology, U.S. Department of Commerce, Washington, D.C. (2023). https://doi.org/10.6028/NIST.AI.100-1
13. OECD: Common guideposts to promote interoperability in AI risk management. OECD Artif. Intell. Pap. (2023). https://doi.org/10.1787/ba602d18-en
14. Artificial Intelligence, H.-L.E.G.: The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment. Guidelines, European Commission, Directorate-General for Communications Networks Content & Technology (2020). https://doi.org/10.2759/002360
15. Aniek, F., et al.: The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Inf. **113**, 103655 (2021). https://doi.org/10.1016/j.jbi.2020.103655
16. Soler G.J., et al.: Harmonised standards for the European AI act. Technical report, European Commission, Seville (Spain) (2024). https://publications.jrc.ec.europa.eu/repository/handle/JRC139430
17. ISO: Information Technology– Artifical Intelligence– Bias in AI Systems and Aided Decision Making– ISO/EC– TR 24027:2021 (E). Standard, International Organization for Standardization, Geneva, CH (2021)
18. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Trans. Inf. Syst. **14**, 330 (1996)
19. Schwartz R., et al.: Towards a standard for identifying and managing bias in artificial intelligence. Special publication 1270, National Institute of Standards and Technology, U.S. Department of Commerce, Washington, D.C. (2022). https://doi.org/10.6028/NIST.SP.1270
20. Niemann, H.: Unter der bank lesen sie alle popper. kritische bemerkungen zu einem artikel von wolfgang welsch (2000)
21. Stein, M., Connor, D.: Learnings from the fda's model of life sciences oversight for foundation models. Technical report, Ada Lovelace Institute (2023). https://www.adalovelaceinstitute.org/report/safe-before-sale/
22. Hicks, S.A., et al.: On evaluation metrics for medical applications of artificial intelligence. Sci. Rep. **12**(1), 5979 (2022). https://doi.org/10.1038/s41598-022-09954-8
23. Jussi, T., et al.: Evaluation of machine learning algorithms for health and wellness applications: a tutorial. Comput. Biol. Med. **132**, 104324 (2021). https://doi.org/10.1016/j.compbiomed.2021.104324
24. Maleki, F., et al.: Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls. Radiol. Artif. Intell. **5**(1), 220028 (2022)
25. Song, Q.C., et al.: Making sense of model generalizability: a tutorial on cross-validation in r and shiny. Adv. Methods Pract. Psychol. Sci. **4**(1), 2515245920947067 (2021). https://doi.org/10.1177/2515245920947067
26. Connor, S., et al.: Adaptability of AI for safety evaluation in regulatory science: a case study of drug-induced liver injury. Front. Artif. Intell. **5**, 1034631 (2022). https://doi.org/10.3389/frai.2022.1034631
27. Austin, C.P.: Opportunities and challenges in translational science. Clin. Transl. Sci. **14**(5), 1629–1647 (2021)
28. Haverkamp, W., et al.: EKG-Diagnostik mithilfe künstlicher intelligenz: aktueller stand und zukünftige perspektiven– teil 1. Herzschrittmachertherapie + Elektrophysiologie **33**(2), 232–240 (2022)
29. Haverkamp, W., et al.: EKG-Diagnostik mit hilfe künstlicher intelligenz: aktueller stand und zukünftige perspektiven - teil 2. Herzschrittmachertherapie + Elektrophysiologie **33**(3), 305–311 (2022)
30. Henry, K.E., et al.: Human-machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. npj Digit. Med. **5**(1), 97 (2022). https://doi.org/10.1038/s41746-022-00597-7
31. Brajovic, D., et al.: Model reporting for certifiable AI: a proposal from merging EU regulation into AI development (2023) arXiv:2307.11525 [cs.AI]
32. Reddy, S., et al.: Evaluation framework to guide implementation of AI systems into healthcare settings. BMJ Health Care Inform. (2021). https://doi.org/10.1136/bmjhci-2021-100444
33. Brocke, J., et al.: Introduction to design science research. In: Brocke, J., et al. (eds.) Design Science Research. Cases. Progress in IS, pp. 1–13. Springer, AG Switzerland (2020). https://doi.org/10.1007/978-3-030-46781-4
34. Joerg, S., et al.: Medaicine: a pilot project on the social and ethical aspects of AI in medical imaging. In: Stephanidis, C.A., et al. (eds.) HCI International 2023 Posters, pp. 455–462. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-35989-7_58
35. Peters, D., et al.: Responsible AI-two frameworks for ethical design practice. IEEE Trans. Technol. Soc. **1**(1), 34–47 (2020). https://doi.org/10.1109/TTS.2020.2974991
36. Palmiotto, F.: When is a decision automated? A taxonomy for a fundamental rights analysis. German Law J. **25**(2), 210–236 (2024). https://doi.org/10.1017/glj.2023.112