AI and Ethics
https://doi.org/10.1007/s43681-025-00723-7

RESEARCH

Implementing AI ethics: the VPCIO model

Désirée Martin[1] · **Michael W. Schmidt**[1] · **Rafaela Hillerbrand**[1]

Received: 30 December 2024 / Accepted: 25 March 2025

Abstract
Due to the development and use of artificial intelligence (AI) systems, there is a need for normative guidance on AI technology. Building on reasonably shared and systematized ethical values and principles (Martin et al. in Comparing AI ethics and AI regulation: ethical values and principles and the case of well-being, beneficence and sustainability, In: Müller, Dung, Dewey, Löhr (Eds.) Philosophy of artificial intelligence: the state of art, synthese library, Springer, Berlin, forthcoming), we aim to provide a framework for implementing ethics in AI systems. The research question in this paper is how to transfer values and principles to an AI system in a way that is understandable and evaluable for users, stakeholders, or an oversight body. Therefore, we work out how to translate values and principles into more concrete norms that can be implemented by the developer and monitored by the executive. Based on our systematization, we extend the so-called VCIO model, where VCIO stands for values, criteria, indicators and observables, as presented by Hallensleben et al. (From principles to practice—an interdisciplinary framework to operationalise AI ethics. VDE, Bertelsmann Stiftung, Frankfurt a. M./Gütersloh. https://www.ai-ethics-impact.org, 2020). Our contribution includes modifications to the model and, most importantly, the addition of principles. Building on this methodology, we present a model that is highly acceptable, the VPCIO model. We developed and evaluated the VPCIO for two case studies. The main case study is an AI-assisted robot used for reconnaissance of radiological hazards (based on a BMBF funded project, entitled KIARA (https://www.itas.kit.edu/english/projects_hill22_kiara.php)). The second case study is about an AI system in an entertaining context, namely to swap faces. Implementing the ethical aspects in these cases into the VPCIO model results in an indicator system that illustrates how ethical aspects can be transferred to an AI system in an understandable way.

Keywords Ethical values and principles · Artificial intelligence · Technology assessment · Responsible research and innovation · AI ethics · Robot ethics · Indicator system

1 Introduction

Over the past few decades, the research, development, and application of AI systems has expanded. The Institute of Electrical and Electronics Engineers (IEEE) explicates in their "position statement" [16] that "artificial intelligence involves computational technologies that are inspired by—but typically operate differently from—the way people and other biological organisms sense, learn, reason, and take action." [Ibid., p. 1].

AI technologies have the potential to improve lives in both personal and professional contexts. However, this development creates not only opportunities, but also risks. It may be important for society to maximize the benefits and minimize the negative impacts, but even if there are no negative impacts, there may be *unethical decisions and actions*. For example, someone could be a victim of bias by being filtered based on gender in the job application process, without being affected by the system's decision because the person was already no longer interested in the job. In AI ethics, an evaluation of (possible) positive and negative consequences is not enough to include all the ethically relevant aspects. For this reason, a deontic and teleological approach is of ethical interest. Thus, there is a need for AI ethics, or more precisely, an implementation of ethics in AI.

✉ Désirée Martin
desiree.martin2@kit.edu

Michael W. Schmidt
michael.schmidt@kit.edu

Rafaela Hillerbrand
rafaela.hillerbrand@kit.edu

[1] Karlsruhe Institute of Technology, Karlsruhe, Germany

Published online: 26 May 2025

In this paper, a reasonable way to do this is to provide concrete ethical requirements for the developer[1] of an AI system, the fulfillment of which will result in an AI system that can be ethically acceptable, at least under ideal conditions.

The high-level expert group on artificial intelligence (HLEG AI) declares in "Ethics Guidelines on Trustworthy AI" (2019) [11], that "[g]lobal solutions are therefore required for the global opportunities and challenges that AI systems bring forth. We therefore encourage all stakeholders to work towards a global framework for trustworthy AI, building international consensus while promoting and upholding our fundamental rights-based approach." [Ibid., p. 5].

To achieve the grand goal of protecting reasonably shared ethical aspects,[2] these aspects need to be considered in the development and design of AI systems. In this paper, the ethical aspects are ethical values and principles.[3], [4]These ethical aspects need to be embedded throughout the life cycle of the system by providing guidelines, standards, and norms. The term 'lifecycle' ranges from the initial idea to deployment, from each updated version to decommissioning. The field of AI ethics is confronted with a number of challenges, including those pertaining to risks and uncertainties, namely the occurrence of negative impacts, e.g. on aspects of safety, security or human dignity. AI ethics is essential to address the challenges posed by the development and use of AI. Challenges in the lifecycle of an AI system can be solved with AI ethics, but we need AI regulations to guarantee them.

For these reasons, many authors, e.g. from organizations and academic institutes, have written normative documents [1, 2, 11, 15]. The 2021 journal article published by IEEE "AI Ethics in the Public, Private, and NGO sectors: a review of a global document collection" [17] affirms that these "documents include principles, frameworks, and policy strategies that articulate the ethical concerns, priorities, and associated strategies of leading organizations and governments around the world." [Ibid., p. 31].

In light of this, a large part of AI ethics strives to embed ethical values and principles into AI technology, or at least provide an approach to what such an implementation might look like (e.g. [2, p. 5]). Such ethical values and principles can revolve around privacy, non-discrimination, responsibility, trust, security, public welfare, sustainability, etc. Based on our comparative study [3], which compared prominent legal frameworks, conventions and ethical guidelines on AI, we identified a near consensus of ethical values and principles and classified these aspects. Since identifying and classifying ethical values and principles is not enough to embed these aspects into AI technology, the results form the basis of the current paper, which elaborates on how these aspects can be transferred into technology. Therefore, the research question in this paper is how to transfer values and principles to an AI system in a way that is understandable and evaluable for different stakeholders, e.g. users or an oversight body. The required level of understanding varies between these different groups and may also vary within the groups, e.g. for different types of stakeholders. We will elaborate how complex morally-laden concepts, i.e. abstract values and principles, can be translated into more concrete norms that are implementable for the developer and monitorable for the executive. For the successful implementation, it is essential that they are understandable to the developer. Our approach is based on Hallensleben et al. [10] who provide an indicator system with values, criteria, indicators and observables. We will implement common values and principles to this system and elaborate criteria, indicators and observables derived from principles. The results of this paper, an indicator system based on the VPCIO model, also have a legal significance in that they provide a blueprint that allows a legal perspective to operationalize a framework with which it would be possible to observe the fulfillment of ethical aspects. In addition, it has significance for developers by providing design indicators. In general, guidelines for embedding ethics are useful in order to present minimum conditions for the acceptability of AI systems. However, research in AI ethics poses many challenges, three of which we will briefly intoduce. First, ethical guidelines in AI ethics often fail to explain moral agency [1, 2, 6]. In this paper, we assume that we probably cannot ascribe full moral agency to AI systems, e.g. [21, 27]. In respect to that AI systems are not morally responsible in the way as full moral agents and we cannot trust them in the way we trust full moral agents. That is why it is especially important to monitor, evaluate and measure the fulfillment of the ethical aspects. Based on this, the research project of this paper focuses on an external implementation of AI ethics,[5] namely the question of how values and principles can be transferred to an AI system in

[1] Or developers. In order not to disturb the flow of reading, we will stick to the singular, whereby we understand this as an abbreviation and include the plural.

[2] For the HLEG AI, these may be the aspects that lead to trustworthiness.

[3] The terms 'values' and 'principles' are discussed in more detail in subSect. 2.1.

[4] This does not mean that ethical aspects can be understood exclusively as ethical values and principles, but in terms of our comparative study and classification, values and principles are the most abstract concepts. Other concepts, such as criteria, indicators, norms, etc., may have an ethical notion, but since this notion can be understood as related to our identified values and/or principles, we focus on ethical values and principles as the relevant ethical aspects.

[5] An internal implementation would mean that we could train the system to act morally.

a way that can be evaluated by different stakeholders, e.g. users or an oversight body.

Second, even if values and principles for AI ethics are addressed, it is not clear whether they are transferable to AI law, because ethical and legal aspects may differ. There are ethical issues that do not need to be translated into a legal dimension, and vice versa.[6] However, regulations are important to ensure ethical aspects that are also understandable as legal aspects. Therefore, we need to identify these ethical aspects and assess their transferability into regulations. In Martin et al. (forthcoming), [3] we compared AI ethics guidelines, international conventions, and AI regulations [1, 2, 6, 11–14, 22, 23, 25]. One result is a broad consensus on shared ethical aspects, which we have classified and systematized as ethical values and principles. With the overall goal of implementing AI ethics, the contribution of this paper is that we provide a transfer of these ethical aspects that is understandable for legislators. This understanding is necessary for them to be able to guarantee these aspects by law. Guaranteeing requires that an oversight body is able to evaluate the fulfillment of these aspects.

Third, in analogy to the second, even if values and principles for AI ethics are addressed, it is not trivial to translate ethical values and principles into technical requirements. The relevant ethical aspects we have identified are too abstract for developers to implement directly. For example, how should a developer implement a morally-laden concept like justice? What is necessary to reasonably claim that justice is implemented? In this paper, we provide a model that transfers these aspects into more concrete requirements that developers can understand. To provide and adopt this model, it is necessary that these aspects are implementable in a concrete AI system developed for a specific purpose. The main case study in this paper is an AI-assisted robot used for reconnaissance and defense against acute radiological hazards.[7] This step of transferring needs to be considered in AI ethics research in order to enable developers to implement the results of AI ethics research into their systems. Our second case study, which concerns a face swapping AI system, serves to test and evaluate the application of ethical transfer in the context of other systems.

Our approach of transferring values and principles with the VCIO model addresses the second, but even more the third challenge. Hallensleben et al.'s indicator system already presents an idea for embedding AI ethics by illustrating a way to move from abstract aspects to less abstract, understandable, and measurable aspects. Their model moves from values to criteria to indicators and from indicators to observables. We have modified the model in two ways. First, many of the ethical aspects, criteria, indicators, and observables we identified differ from the values, criteria, indicators, and observables presented by Hallensleben et al. For example, according to our findings, we have identified three values: understanding, justice, and well-being. Second, we modify it with an additional category. In particular, their model lacks the category of principles, but we have three reasons for adding it. First, principles are less abstract than values and thus provide a link between values and criteria. In addition, they trace ethically required actions. Second, AI ethics guidelines focus mainly on principles, or at least on aspects that can be conceptualized as principles. Third, based on this literature, some principles can be understood as reasonably shared principles. By modifying Hallensleben et al.'s model and adding principles, we close this gap, take up the AI ethics guidelines and provide a more practical solution of the indicator system.

In practical terms, it is not expected that the systematization can be implemented in the same way in every AI system. Furthermore, it is unlikely that the indicator system would lead to the same indicators regardless of the context of the AI system. This paper argues that even if a blueprint for transferring values and principles is available, it must be adapted and possibly modified to fit specific contexts and AI use cases. To illustrate and test this thesis, we first developed an indicator system model for the main case study. Secondly, the thesis is evaluated and tested with a second case study, namely an AI system in an entertainment context, specifically a face-swapping system.

The following section outlines briefly the VCIO model developed by Hallensleben et al. and identifies its benefits and shortcomings (2.1), presents a conceptualization of values and principles (2.2), and a modification of the model (namely the VPCIO model) that incorporates AI ethics guidelines. The advantage of this modification is that it provides a meaningful shared and systematized set of ethical values and principles (2.3). The last part of Sect. 2 illustrates our concretized results of values, principles, and criteria (2.4). The third section deals with the evaluation of the fulfillment of ethical aspects in AI systems by showing the potential applications of the model. The conclusion provides a short summary and answers the research question.

[6] For example, it may be ethically unacceptable to cheat on your partner without it being a criminal offence. At the same time, it may be legally important to enact a traffic rule that says right before left, without this rule being ethically necessary.

[7] The project, entitled KIARA, is funded by the Federal Ministry of Education and Research (BMBF). https://www.itas.kit.edu/english/projects_hill22_kiara.php

2 The VPCIO model

2.1 From VCIO to VPCIO

In their article "From Principles to Practice. An interdisciplinary framework to operationalise AI ethics"[10] by S. Hallensleben et al., the authors present the VCIO model (Values, Criteria, Indicators, Observables) as "an approach for the […] operationalisation of values" (ibd., p. 8). The benefits of this model are ethical aspects, that are "practicable, comparable, and measurable" (ibd.). The VCIO model has already been revised and refined [9]. In both versions, observables are categories of possible responses to indicators phrased as questions. These questions are designed to capture specific criteria that specify values. The following Fig. 1 illustrates the model and attempts to reflect both versions.

Hallensleben et al. [10, 9] also provide practical examples of the application of the VCIO model to designated values. The values embedded in the VCIO model in Hallensleben et al. [10, pp. 21–25] are 'transparency', 'justice' and 'accountability'; the values in Hallensleben et al. [9, p. 8] are 'transparency', 'accountability', 'privacy', 'fairness' and 'reliability'. We consider this model to be highly applicable and easy to understand. It presents values that are as such highly abstract, but that are understandable as reasonable ethical aspects. The criteria define when values are fulfilled or violated. As an example, the criterion "Documentation about the AI systems operation" [9, p. 9] defines when the value of transparency is fulfilled or violated [ibid.]. These criteria may not be unequivocal ethical concepts, but they are less abstract than the clearly ethical values and they refer directly to them. Indicators like "Are the characteristics of the AI system(s) documented?" [ibid, p. 11] monitor whether the criterion like "documentation about the AI systems operation" (s.a.) is met. The question format of the indicators allows for different levels of possible answers, namely the observables. Hallensleben et al. differentiate between seven possible answer levels, so that the ethical evaluation of a system has the potential to be rich in detail, traceable and understandable. According to our understanding, the approach of the VCIO model is a practicable way to compare and measure ethical aspects in AI systems. This view also corresponds to their view in Hallensleben et al. [10, p. 8, see above].

However, contrary to what might be expected from the title, Hallensleben et al. do not state any principles, or at least their understanding of principles remains unclear. They state that they propose the VCIO framework "to bring ethical principles into actionable practice when designing, implementing and evaluating AI systems" [10, p. 6]. In another formulation, it says that they "present the VCIO model (values, criteria, indicators, and observables) for the operationalisation and measurement of otherwise abstract principles" [ibid] and that they introduce "a framework that

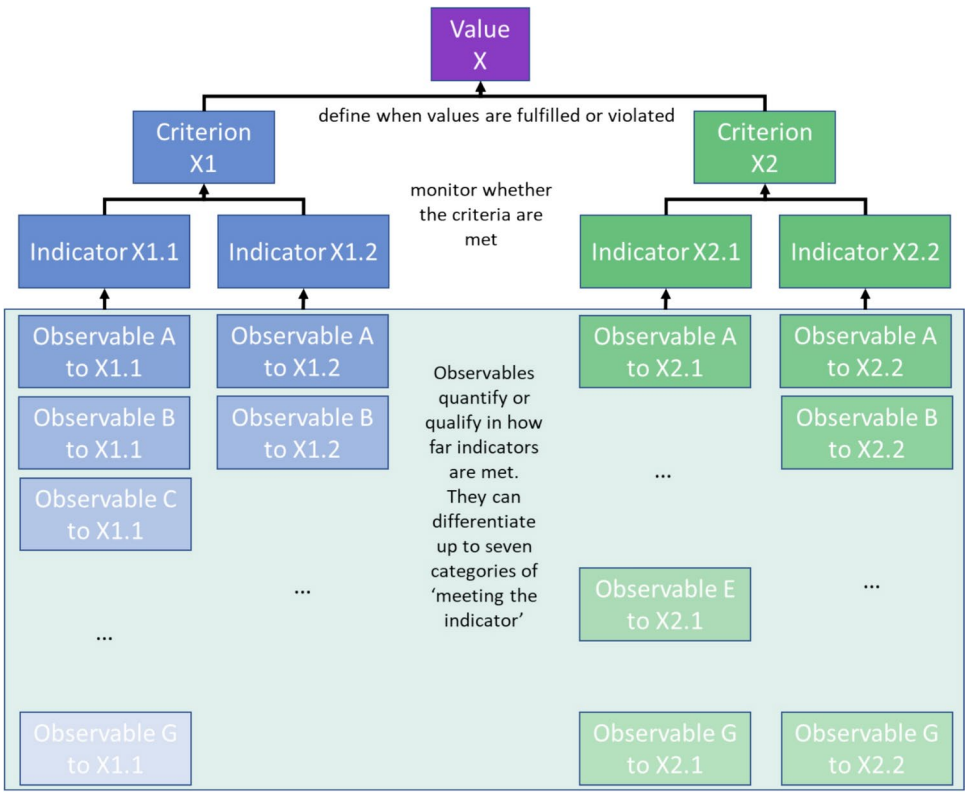


Fig. 1 The VCIO model in 2020 and 2022. See Source: Fig. 2 in [10, p. 16] and Fig. 1 in [9, p. 7]

demonstrates how to put ethical principles into AI practice" [ibid., p. 8].

Indeed, they write in one sentence "taken together, our approach for the operationalisation of general principles (VCIO) […] provides a framework for bringing AI ethics from principles to practice" [10, p. 14] and this might sound like that they have a broad understanding of principles and that values, criteria, indicators and observables are part of them. Another interpretation would be that the aspects of the model (VCIO) refer to ethical principles that are not named. This theory would be supported by the fact that they speak twice of "principles and values" [ibid., p. 8, 26], so it seems that they understand values not included in principles.

Our approach is that principles are important for AI ethics, but they do not seem to be explicit in the VCIO model. In our understanding, principles are action-guiding, so they have the potential to dictate and determine what the AI system or the AI developer should 'do'. The VCIO model already presents hierarchical levels from abstract values to concrete indicators, but since values are vague, abstract, and do not guide action, we need a conceptual bridge from values to criteria. Therefore, we are convinced that an AI ethics framework is even more practicable and the fulfillment of ethical aspects even more measurable, when principles are part of the framework. This is a first reason for us to modify the VCIO model by adding principles.

It is also worth taking a look at the literature. Various prominent AI ethics guidelines present specific principles [2, 6, 11, 13], so our approach of including specifically named principles is in line with the literature in AI ethics. This is a second reason for including principles.

The last but not least reason for including specific principles is the principles in the context of AI ethics. Looking again at the literature, various ethically important principles are presented, and many of them are shared principles in the different documents, e.g. (respect for (human)) autonomy [2, 6, 11].

This insight leads us to do more than add principles as important aspects of AI ethics to the VCIO model. We present a modified model based on shared ethical aspects with respect to the AI ethics guidelines. Our goal is to provide a framework, called the VPCIO model, that is highly acceptable from a practical and ethical point of view.

The VPCIO model and its specifications based on the case studies are presented in the following section.

2.2 Conceptualization and the VPCIO model

Our VPCIO model is based on the VCIO model and mainly on the refined version of the Hallensleben et al. [10] indicator system. With the goal of an indicator system that incorporates shared values and principles, we have added principles. In order to include principles in the model, it is necessary to consider the concept of values and principles. Otherwise, it is neither understandable why the principles are added in a certain level nor reasonable from the perspective of ethics research. Based on the explications of the ethics guidelines [11, p. 22] and other prominent notions [20, 4], our approach is that values can be understood as desirable states [ibid.]. In this sense, a value can reasonably be classified as a final social good or as socially desirable, so that values refer to goodness, "or what we believe [epistemic dimension] [...] or express [semantic dimension] [...] to be good" [28, p. 302]. This understanding may also be the reason why Hallensleben et al. present these most abstract ethical aspects, as they "form the basis of AI ethics and are open to different interpretations" [10, p. 10].

Principles[8], on the other hand, refer to an action-guiding norm according to which one should act from a moral perspective [18]. In his Critique of Practical Reason, Kant refers to principles as laws and maxims, and also mentions rules derived from them. [ibid.] These concepts illustrate his idea of action-guiding principles. In this sense, principles do not (directly) refer to what is good, but rather to what is right.

"'Right' and 'good' are [therefore] the two basic terms of moral evaluation. In general, something is 'right' if it is morally obligatory, whereas it is morally 'good' if it is worth having or doing and enhances the life of those who possess it" [20]. With that in mind, and also in reflection to Ross 1930, we attribute *rightness* to actions (principles) and *goodness* to persons and things (values).

In terms of this conceptualization, modified systematization adds principles between values and criteria, so that principles can also be understood as links between values and criteria. Systematization means identifying relationships between ethical aspects, especially between values and principles, and interrelationships between principles. The relations between values and principles are interrelations and each principle is related to at least one value. However, the category 'principle' is not sufficiently granular: principles can be classified into higher- and lower-levels. (Fig. 2) These principles have 'intra-relations' because the relations are within the different principles. The difference between the levels is typically the level of abstraction. The higher-level principles and values are highly abstract and these principles are in line with Beauchamp and Childress' mid-level principles [4, 7], that have long been shaping the direction of biomedical ethics. The lower-level principles

[8] In the history of philosophy, the concept of principle has been used in different fields and with different meanings. In this paper, we will refer only to principles in the field of morality. We could have called these principles *moral principles*. However, since we are only referring to this type of principle, we will stick to the term *principles*.

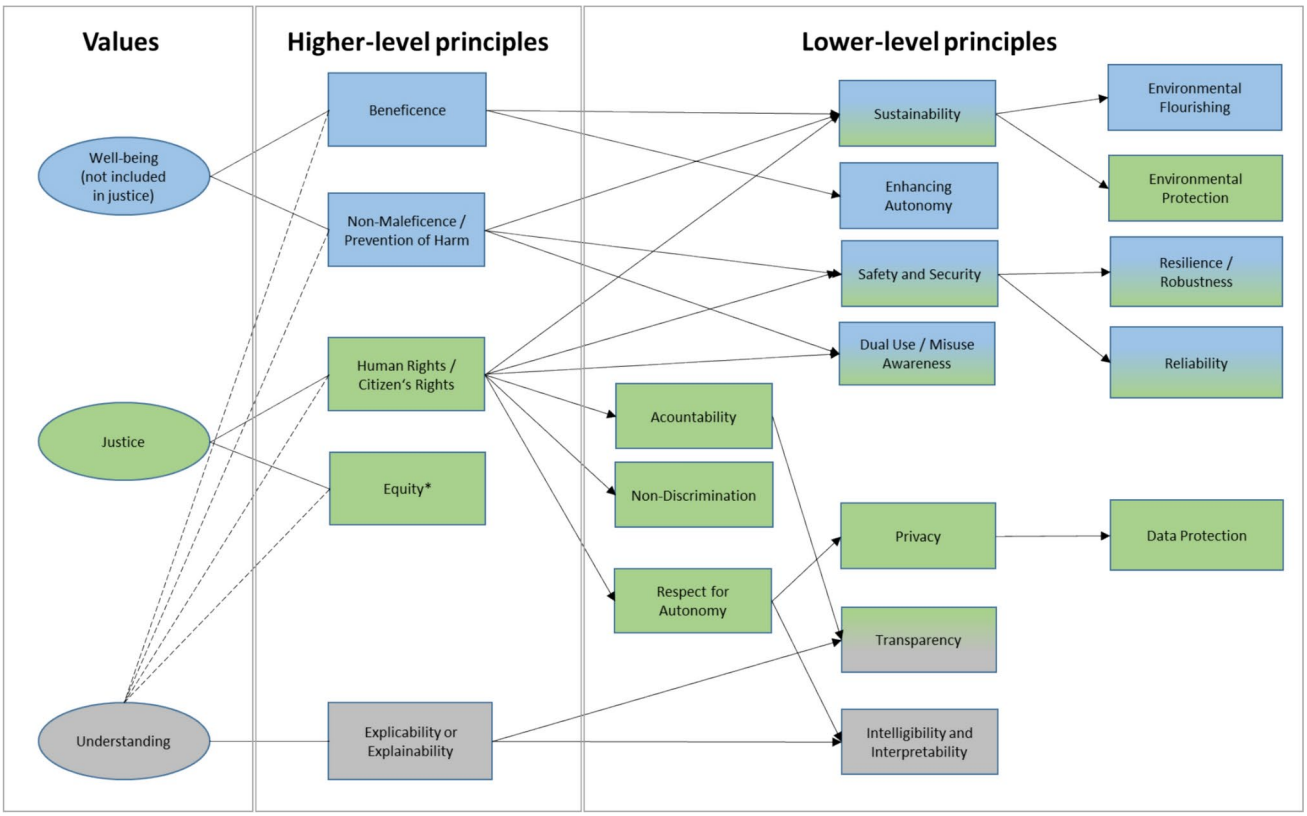


Fig. 2 Systematization of values and principles Source: [3]. * Equity is not part of the consensual comparison, but in our understanding, it is an important ethical aspect and therefore part of our systematization (Relationships are not necessarily limited to those shown here. Different values and principles can influence each other, in particular the value of'understanding' can influence all higher-level principles, because we do not imagine a principle of'beneficence', for example, to be carried out without including considerations of understanding. Some principles can be understood as being related to two values, and so they are shown in two colors. Our values of well-being, justice, and understanding contrast with those of Hallensleben et al. [9]. The values of well-being and justice broadly understood have significant overlap, but for our systematization we propose an interpretation of well-being that goes beyond aspects of justice because it requires more than what is derived from aspects of justice.)

are typically more specified with respect to specific contexts. The lower-level principles, which are very close to the higher-level principles (see Fig. 2), probably form an intermediate point here. For example, the principle 'respect for autonomy' is a lower-level principle that we have assigned to the higher-level principle 'human rights'. In the context of biomedical ethics, however, it is traditionally understood as a mid-level principle [4]. This is also conceivable here, e.g. in AI contexts that could have a particularly serious impact on human autonomy, 'respect for autonomy' could play a more prominent role.

This is particularly important because specified requirements promote the traceability of the transfer of AI ethics to a specific AI system.

Criteria indicate whether a principle is performed or not by providing ways in which the principle can be performed. Let us consider a possible principle of misuse awareness. A related criterion concretizes how to perform this principle, so a possible criterion might be the avoidance of misuse. The criteria are therefore more concrete than the principles, and as such often formulated in more than one word. Noting that not every higher-level principle is completely reducible to lower-level principles, higher-level principles may also be directly related to criteria. Taking this into account, criteria are one step closer to specific requirements. The next step is to develop indicators to monitor whether the criterion is being met. Indicators are formulated as questions. These questions can be used to ask about different characteristics that need to be considered in order to meet the related criterion. Finally, the observables provide different answers to an indicator.

The following Table 1 illustrates our conceptualization, important relations of these concepts and provides examples of them.

As indicated by the arrows in the "relationship" and "level of abstraction" columns, the VPCIO model is primarily hierarchical, with the exception of values and principles. Values and principles are hierarchically higher than criteria, which are higher than indicators, which are higher than

Table 1 The concepts of the VPCIO model and their relations

Level of abstraction[a]	Concept		Short explication		Example	Relations[b,c]		
Abstract	Values		Desirable state, an ethical good		Well-being		interrelations and each principle is related to at least one value	↓↑
	Principles	Higher-level principles	Action-guiding, refer to what is right	In line with the mid-level principles (see above)	Beneficence	intrarelations between higher and lower-level principles		
		Lower-level principles		Typically more specified	Resilience / Robustness			
	Criteria		Demanding formulation		Robustness and Reliability in Operation[d]	concretize how to perform a related principle		↑
	Indicators		They are formulated as questions to allow for a clear answer.		Is the applied AI lifecycle management robust to changes in the operational domain?	monitor whether the criterion is being met		↑
Concrete	Observables		Up to 7 possible answers of the indicator question		Yes	Aim to quantify whether or to what extent an indicator is being met by answering the related indicator question		↑

[a]This includes also indicators and observables, as the observables answer and therefore specify what is being asked by the associated indicator

[b]The relationships are not necessarily limited to the ones mentioned, but the ones mentioned are relevant for this paper because these concepts and their relationships allow for an indicator system for AI ethics

[c]The arrows illustrate the direction of the relationship

[d]The example of a criterion and an indicator is adapted from Hallensleben [9, p. 41]

observables.[9] However, values and higher-level principles are not necessarily hierarchically ordered because they can refer to different ethical theories. With regard to the question whether the good (the values) are prior to the right (the principles) or the other way around, we stay therefore agnostic and only assume some sort of close relation.

With this fine-grained system of indicators and observables, the VPCIO model is adaptable to different AI systems. While all of these systematized ethical aspects are elaborated with respect to our main case study, which concerns an AI-assisted

[9] In addition, the higher-level principles are hierarchically higher than the lower-level principles.

reconnaissance robot, we test their applicability for other AI systems with the second case study, which concerns an AI system developed to swap faces. It is essential to evaluate the applicability of these ethical aspects, as it is not realistic to expect that the indicator system (or any other model) can be implemented in the same way in every AI system.

Let us consider a possible principle in AI ethics such as 'enhancing autonomy'. We can relate this principle to a value called 'well-being'[10] and the criteria for this principle might be 'enhancing human decision making', 'Enhancing physical human capabilities' and 'Enhancing human creativity and individuality'. If a system enhances physical human decision making, their physical capabilities and their creativity and individuality, we can reasonably say that the system *is acting* in accordance with the principle of enhancing autonomy.

With these explanations in mind, we approach how our provided VPCIO model aims at implementing shared ethical aspects in an AI ethics practice. This model could be (part of) an answer to the research question of how to transfer values and principles to an AI system in a way that is understandable and evaluable for different stakeholders, e.g. users or an oversight body. The following chapter illustrates the way to systematize values and principles as a first step to implement values and principles in the VPCIO model.

2.3 Systematizing shared ethical aspects and the role of ethics in AI regulations

In Martin et al. (forthcoming), we compared prominent documents with academical background [3].

From an ethical perspective, the widely discussed guidelines compared [1, 2, 6, 11, 14, 15] share an ethical perspective on AI but they differ, for example, in the ethical values and principles that they cite and in their terminology. Summarizing the results of that paper, there is a near consensus on the aspects used in the ethical guidelines. 'Accountability' or 'responsibility', various aspects of 'autonomy', 'well-being', 'transparency' and 'explicability' are ethical aspects that are considered in almost every considered guideline. They also refer to aspects of 'justice' and 'fairness', 'human rights' and 'human values', aspects of 'non-maleficence', and 'privacy'.

Based on the classification of values and principles and the consensus on ethical values, it was possible to systematize the ethical aspects. The systematization of the respective aspects and the relations or connections of the respective aspects are illustrated in the following Fig. 2.

Let us reconsider our conceptualization of principles as something that guides action, something that refers to what is right rather than what is good. This means that the principles presented in Fig. 2 have such an action-guiding and rights-referencing component. Human rights may not be directly action guiding and principles as defined by Ross. However, as we are considering a specific context, namely AI-assisted robots in the scenario mentioned, they translate into action-guiding principles in a more or less straightforward way. This might be best seen with Article 2 in the Declaration of Human Rights, which contains formulations of its articles such as "Everyone has the right to […]" [26] Article 3, 6, 8, etc. "No one shall be" [26] Art. 4, 5, 9, 12, etc.[11,12], but also for example when we think of article 1 "All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood", we can reasonably argue for intergenerational justice which in turn morally obliges us to perform the principle of sustainability.

In addition, we compared these ethical guidelines with prominent international conventions [22, 25] and legal frameworks [12], particularly at the European level [23], the first and most recently enacted AI regulation at that level. This comparison is important to assess possible consensus and disagreement, which is necessary to develop recommendations for change [3].

The comparison produced two results, a consensus and a dissensus. First, the dissensus: 'well-being' and two related principles seem to be considered obligatory in many documents [2, 6, 13, 22], but in the EU AI Act they do not seem to be obligatory, or at least are not operationalized as such [23]. If ethical considerations are to be transferred to the legal dimension, research is needed to explain why ethical considerations aren't as clearly operationalized in the legal realm as they are in ethics.

In terms of the current paper, the second result may be of greater interest: There is almost consensus. Examples of shared aspects in almost all ethical and legal documents are 'accountability' (or 'responsibility'), various aspects of 'autonomy', 'well-being', and 'transparency' (or 'explicability'[13]). The differences in the conceptualizations are

[10] In some contexts, autonomy does not seem to be directly connected to well-being. For example, a suicide can be a result of an autonomous choice, but does it foster well-being? According to our interpretation, well-being can entail individual reasonable desirable states. In our example, the autonomous decision to end one's life prematurely with a terminal illness can bring the decision-maker a peace that contributes to their well-being.

[11] Accordingly, the other principles are considered here, e.g. the principle of safety and security might be something like "Everyone should be safe and it's not allowed to impede anyone's safety".

[12] Further information on our higher-level principles can be found in Beauchamp and Childress [4], which, for example, presents a principle of beneficence and non-maleficence for the context of biomedical ethics.

[13] 'Interpretability' may be of practical importance to the developer of an AI system, and may be more important to them than explicability or explainability. However, our systematization aims at AI ethics, so

mainly in their functions and importance (e.g. that is the case for ‘well-being’).

With these results, our systematization can be understood as a first step towards an AI ethics recommendation that also discloses legal interests in the ethical aspects, which is important to ensure AI ethics.

These results are therefore important for the research question of how to transfer values and principles to an AI system in a way that is understandable and evaluable for different stakeholders, e.g. users or an oversight body[14]. Since regulations are necessary to ensure ethics, it is crucial to see where regulations include ethical aspects and where they are missing. However, this is not enough to promote understanding of the transfer of ethical aspects for an oversight body that could be part of ensuring ethics. An oversight body needs at least a high level of understanding of the necessary observables[15] in order to evaluate their fulfillment in a use case of a specific AI system. Based on this evaluation, the AI system can be labeled with the AI ethics label for the tested use case. This label is useful for the user because it clarifies the system's connection to AI ethics. Therefore, this is part of the answer to our research question regarding the transfer of values and principles to an AI system in a way that is understandable for the user. Even when the user does not know about the observables (due to contingent limitations like time), the user should be able to understand whether and to what extent the system fulfills ethical values.

The problem at this point is how the systematization of values and principles can lead us to an application in practice. The values and principles are abstract, and it is not clear or unequivocal how they should be interpreted and transferred to AI ethics in practice. However, this is necessary to answer the research question how to transfer values and principles to an AI system in a way that is understandable and evaluable for different stakeholders, e.g. users or an oversight body.

2.4 The VPCIO model

In order to approach the transfer of systematized and shared ethical values, we focus on the idea of an indicator system, namely the VPCIO model. In order to understand the reasons for considering the VCIO model of Hallensleben et al. and the need to modify it, we compare the VCIO and VPCIO models. In addition, we provide a broad picture of the concretized VPCIO model, in a way that is understandable and evaluable for different stakeholders.

Looking more closely at the precise systematization of their VCIO model and our VPCIO model[16], we can see some commonalities, but also some differences. The already mentioned values of Hallensleben et al. [9]—'transparency', 'accountability', 'privacy' and 'reliability'—are also part of our systematization, but we have classified them as principles. According to our understanding, 'justice' is understood as a value[17]. This is a commonality with Hallensleben et al. [10]. In the refined version, Hallensleben et al. [9] modified this value and replaced it with the value 'fairness'. However, based on our comparison of AI ethics guidelines, conceptualization, and consideration of AI regulation, we retain the value of 'justice' because it is a broader concept that includes the principle of 'human rights/civil rights' and that of 'equity'[18]. However, based on Hallensleben et al. [10, 9], their aspects of fairness and justice are considered and included in the VPCIO model. Especially for the other two values, 'well-being' and 'understanding'[19], we developed new criteria for the designated principles related to these values, but also for some aspects of ‘justice’.

Taking a closer look on our modified VPCIO model, criteria are not necessarily directly related to lower-level principles. It is possible, that principles are not fully reducible to lower-level principles. The development of the specific

the focus is on the use of an AI system and the directly or indirectly affected stakeholders, such as users or society. Taking this (and AI ethics guidelines) into account, we classify 'explicability and explainability' as an overarching principle that includes aspects of transparency and aspects of interpretability and intelligibility (Fig. 2).

[14] More details on different stakeholders in Sect. 3.3.

[15] Assuming that negative anchor indicators, etc. are defined by another body.

[16] The model is not complete. Some criteria, indicators and observables have been taken over from Hallensleben 2020 and 2022 [10], [9], others have been refined and some are new. However, especially at the level of observables, the list is not exhaustive, but sufficient to present, adapt and evaluate the model in relation to this scope and the two case studies.

[17] Since we have recognized principles as something that refers to what is right rather than what is good, justice might be seen as a special case. Justice can be understood as a principle, as it can indicate what is right, but it can also be understood as a goal, or at least a just state can be understood as a good, a goal and a value. In this paper we share a very broad notion of justice, which can be understood as something desirable and therefore as a value, and which refers to various principles that support its realization.

[18] We understand ‘justice’ as a value according to Hallensleben et al. [10] and IEEE [13]. Justice is also an important morally-laden concept in five out of six ethics guidelines. In contrast, 'fairness' is highlighted in three out of six ethics guidelines. In this paper, we understand fairness to be implicitly included in the value of justice and related principles, such as the principle of equity.

[19] In the context of AI, understanding is seen as important. Even if it is not a classical ethical value, it can be understood (at least in this context) as an epistemic value or an epistemic good [8], and therefore as something we *desire* to achieve. An epistemic value that is connected to the morally good as realizing ethical principles (or principles related to ethical values) often seems to require a detailed understanding of the algorithm. It is therefore not just a means to an end and can be seen as a desirable state. We also retain the notion of understanding as a value associated with a shared principle of explicability or explainability.

criteria is not only based on a nearly consensus of principles, but also of ethically important aspects considering our case study of an AI-assisted reconnaissance robot. These aspects affected the specific principles, but also the criteria, indicators and observables. In the following passages, we focus on the values, principles and criteria. Afterwards, we provide in the next chapter an example of indicators and observables that illustrate how our values, principles and criteria lead to less abstract and instead more assessable requirements. Additionally, the evaluation of the applicability of this model to other AI systems, specifically one for swapping faces, is discussed.

2.4.1 The values, principles and criteria in our VPCIO-model

First, the value of'well-being' and their directly related principles and criteria (Fig. 3).

Our VPCIO includes new criteria for 'dual use/misuse awareness','environmental flourishing','enhancing autonomy'. These new criteria are 'DU/MA1 Avoidance of Misuse'[20] for the principle 'dual use/misuse awareness', the two criteria for the principle 'environmental flourishing', namely 'EF1 Life cycle assessment', 'EF2 AI systems decisions enhance environmental flourishing and sustainability' and the three criteria for the principle 'enhancing autonomy', namely 'EA1 Enhancing human decision making', 'EA2 Enhancing physical human capabilities' and 'EA3 Enhancing human creativity and individuality'. An exception is the principle of safety and security and their lower-level principle 'resilience/robustness/reliability'. The criteria, indicator and observables could be adopted with minor modifications from Hallensleben et al. [9, pp. 41–47][21].

Second, the value of 'justice' and their directly related principles and criteria (Fig. 4).

In our VPCIO the criteria for the principle of 'accountability', 'privacy', 'resilience/robustness/reliability' are taken entirely from their value-labelled aspects 'accountability', 'privacy' and 'reliability' in Hallensleben et al. [9], with the mentioned minor modifications for 'resilience/robustness/reliability' [9, pp. 16, 24, 41]. The principles 'resilience/robustness/reliability' and 'dual use/misuse awareness' are not only related to the value of well-being, but also to the value of justice. This is illustrated by the two-colored signs. In addition, we have adopted from Hallensleben et al. [9] the criteria for 'transparency' T1, T2 and T4, the criteria F1 of 'fairness' for our principle of 'non-discrimination' and the criteria F3 of 'fairness' for our principle of 'environmental protection' [ibid., p. 9, 30]. We adopted also a criterion of 'justice' from Hallensleben et al. [10, p. 23] for our principle of 'respect for autonomy'. However, we developed also new criteria for the principle of 'equity', namely 'E1 impact of just structure of society', for the principle of 'transparency', namely 'T3 uncertainties in the detection process', for the principle of 'respect for autonomy', namely 'RA2 informed consent' and for the principle of 'privacy', namely 'P4 protection of spatial privacy'.

Third, the value of 'understanding and their directly related principles and criteria (Fig. 5).

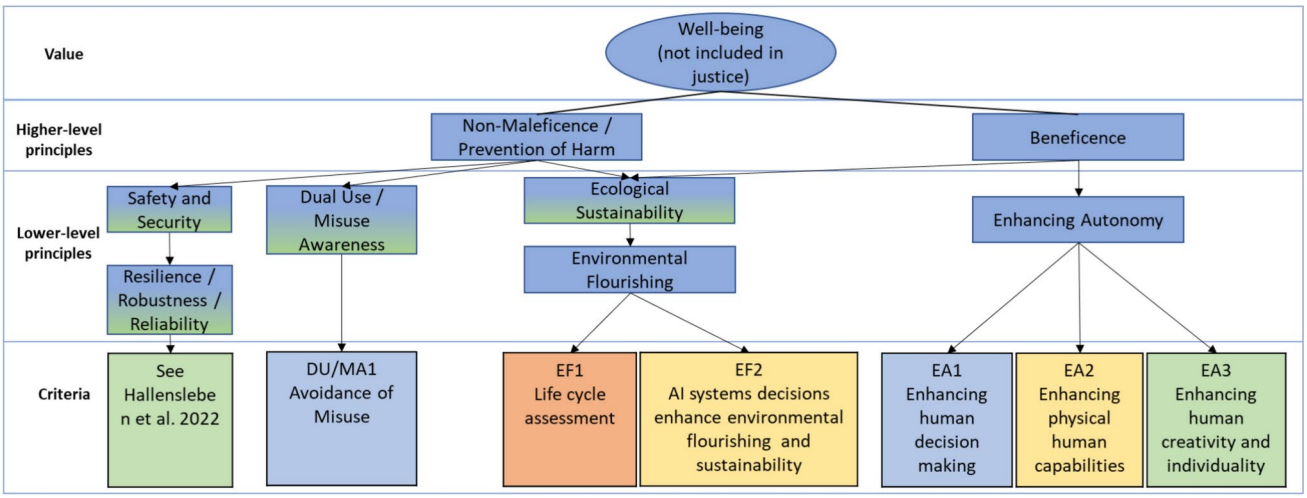


Fig. 3 VPC of well-being

[20] The abbreviations derive from the name of the related principle. In this case, this first criterion, 'DU/MA1' is related to the principle of 'Dual Use / Misuse Awareness'.

[21] One modification is that we have refined the formulation of R1 and R2, as we use resilience as an umbrella term for robustness and reliability.[5] Another modification is that some of the indicators are rewritten to also refer to resilience, which is not explicitly mentioned in the reliability aspects of Hallensleben et al. [9].

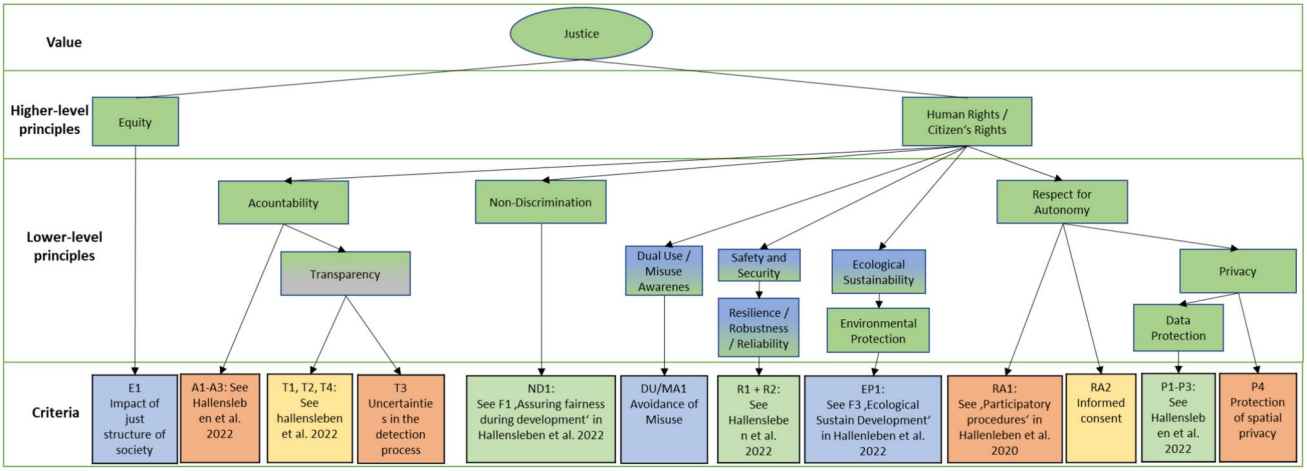
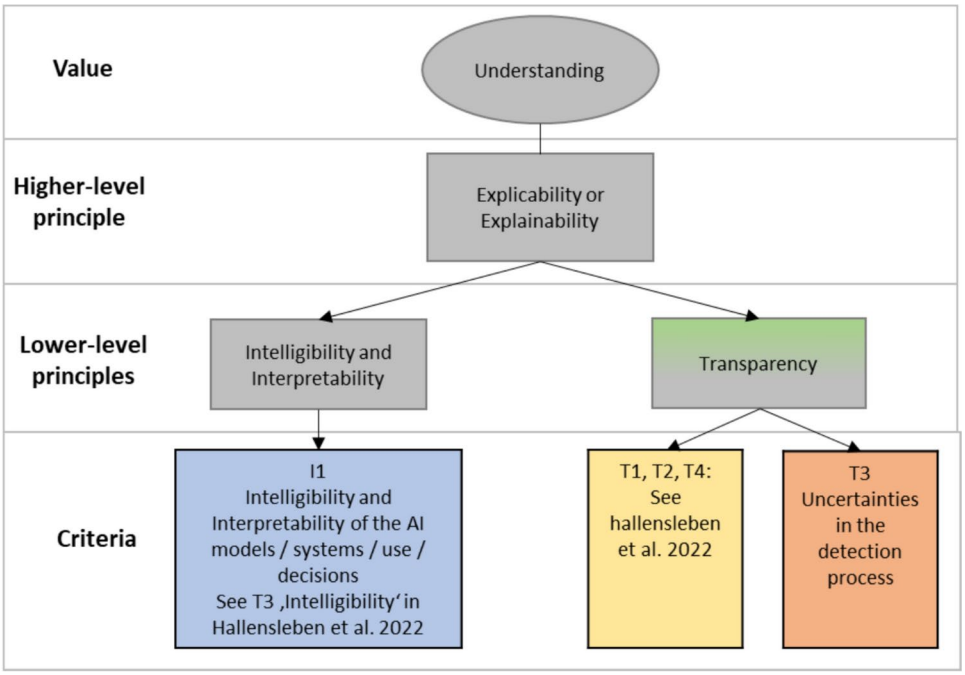


Fig. 4 VPC of justice

Fig. 5 VPC of understanding



The two-colored sign for transparency illustrates that this principle can be understood related to the value of understanding but also of justice. Again, we find the principle of 'transparency' with their criteria form Hallensleben et al. 9 (T1, T2, T4). Additionally, we adopted the criterion 'T3 intelligibility' from Hallensleben et al. [9, p. 9] for our principle 'intelligibility and interpretability' [ibid.].

3 Applying AI ethics to practice

Let us now turn to the importance of case studies. They are important not only for the development of criteria, indicators, and observables, but also for their evaluation and testing for completeness, as well as for the goal of evaluating AI ethics. While the first two points primarily serve to develop an applicable VPCIO model, the third point aims to evaluate a system both prospectively and retrospectively. This means that prospectively possible ethical aspects can be considered during the development of a system, and retrospectively they serve the ethical assessment of the system.

3.1 The first case study and its role in the development of criteria, indicators and observables

The first and main case study in this paper is (based on the KIARA project) an AI-assisted robot used for reconnaissance and defense against acute radiological hazards.

The VPCIO model was built in respect to this case study. In this, the AI system used serves a good purpose and can benefit society by minimizing high health risks for emergency forces and detecting radiological sources that could endanger civilians and the environment. However, even if a system is developed with good intentions, there are still ethical issues to consider. In this particular case study, several possible stakeholders need to be considered: the civilians (and non-human beings) who may have direct contact with the AI-assisted robot, the civilians and the environment who do not have direct contact, the operator of the system, the management of the operation (to name a few). Each party may give rise to different ethical considerations. Consider the following scenario: emergency responders receive an alert about a radiological threat. The emergency forces are arriving near the possible danger zone. The robot is sent into the danger zone first to avoid endangering the emergency personnel. In this scenario, it is important that the operator is well trained to neither underestimate nor overestimate the AI tools. Based on this training, the operator can guide the robot to potentially interesting locations, is able to interpret the possible detection of a hazard symbol or the detection of a human being. In contrast, for civilians with direct contact, the purpose of the robot is especially relevant for them and should be transparent. In this scenario, this would be ethically important for the detected human in the danger zone. However, civilians outside the danger zone may also need some information about the emergency situation, so in terms of the principle of transparency, different parties may need different levels of transparency to understand the information relevant to them.

All these aspects need to be considered in the VPCIO model and in the ethics assessment of an AI system with this model. The assessment is already taken into consideration in Hallensleben et al. [10, 9]. They present a so-called AI ethics label "inspired by the well-known energy efficiency label" [10, p. 12] as part of "a framework for bringing AI ethics from principles to practice" [10, p. 14]. In the AI ethics label each value is assigned a letter (A to G). The range from A to G indicates the degree of value fulfillment [10, p.

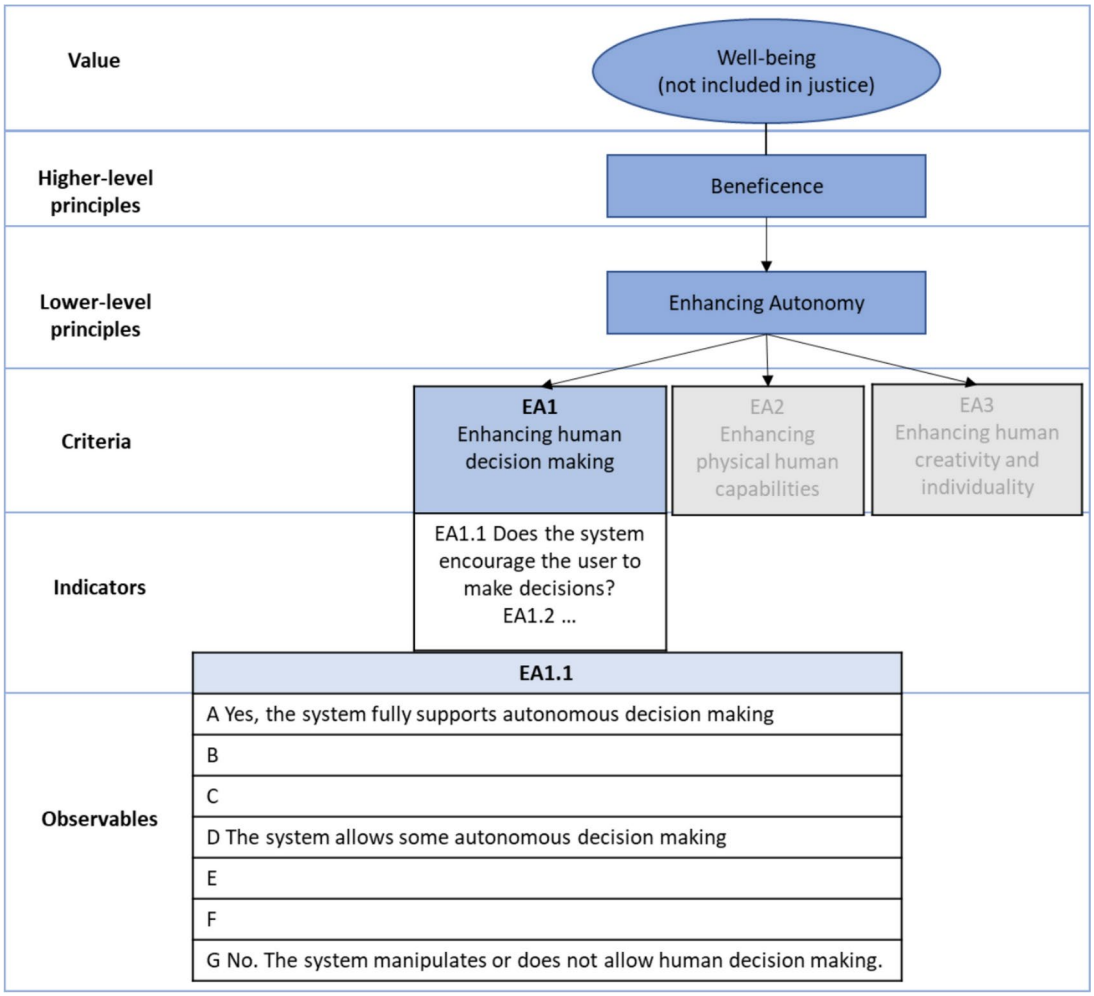


Fig.6 The indicator EA1.1

13, 18, p. 50]. At this point it becomes clear why the possible observables are distinguished into 7 different answers. To illustrate, let us consider the example of the indicator 'Does the system encourage the user to make decisions?' (Fig. 6):

Answering this question with observable A "Yes, the system fully supports autonomous decision making" would mean, in the context of the AI-assisted robot, that the AI system provides all relevant information to the operator so that they are able to make an appropriate decision. In addition, when the AI system suggests decision options, the system does not emphasize or manipulate the operator in any direction. In furthermore, the operator can choose one of the options suggested by the system, but also choose another option as easily as the suggested one.

To assign a letter to the value of well-being, the indicator levels are aggregated to a level of the corresponding criterion. This aggregation process is repeated for the different levels to assess the fulfillment of a value for a particular AI system. The details of the aggregation can be found in chapter 5 of Hallensleben et al. [9, p. 48ff.]. Their aggregation has three special features, i.e. "positive anchor indicators" [ibid., p. 48], "negative anchor indicator" [ibid.] and "skippable indicators" [ibid.]. If the designated positive anchor indicators are fulfilled, the corresponding criterion is fulfilled to the same extent as this indicator. Skippable indicators mark conditions that are not necessary, i.e. they do not need to be fulfilled or "taken into account for aggregation" [ibid.]. In this paper, we would like to highlight the feature of the negative anchor indicators. When such indicators are defined, they "are necessary conditions to meet the aim of a criteria. If a negative [anchor] indicator is not sufficiently fulfilled, the indicator cannot be fulfilled either [and] the whole criterion cannot be fulfilled" [9, p. 48f.]. In Hallensleben et al. [9], a negative anchor indicator is not sufficiently fulfilled if it is marked with a "G", but we think there is another possibility [ibid.]. There may be scenarios in which we would argue that a particular rate of "D" would be ethically unacceptable, so the entire criterion would be rated "D". Furthermore, this specific "D" would mean that the criterion is not sufficiently met. To illustrate this thesis, we will give an fictional example, not based on the KIARA robots (Table 2).

In the specific case of the AI-assisted robot, it may be necessary that some indicators of well-being are minimally rated with a "C".[22] R 1.6 is a negative anchor indicator in this example.

One negative anchor indicator, among several others that are possible but not shown in this table, is R 1.6, which we adopted from Hallensleben et al. [9] with the addition of 'reliability': "Are measures in place to ensure the integrity, robustness, [reliability] and overall security of the AI system/application against potential attacks over its life cycle?" [9, p. 45]. In the case of an AI-assisted reconnaissance robot, the negative anchor indicator is understood as not being met with a "D", which in this case means "Measures partly defined and information can be retrieved by a defined process." [ibid.]. In the context of our special case, i.e. in the area of safety and health aspects relevant to society, it is necessary that at least the measures are completely defined and "can be retrieved by a defined interface" [Observable "C", ibid.].

Since the negative anchor indicator R 1.6 is rated with a "D" and this is already considered "not fulfilled", criterion R 1 is also not met and is rated with a "D". In Hallensleben [9], an unfulfilled negative anchor indicator, which they rate as a "G," means that the criterion is not met. We would like to discuss another option. The aggregation of higher levels depends on the answer to the question of whether negative anchors can or should be extended to higher levels. If not, the lower-level principle "Resilience/Reliability/Robustness" is rated with a "C", so that the higher-level principle "Nonmaleficence/Prevention of harm" is rated with a "B" and the value of well-being is also rated with a "B" in this scenario. If the ethical unacceptability of the indicator not met, namely the negative anchor indicator R 1.6, is understood as a direct influence on the other levels and as something that would mean that the entire value is understood as not met, we would grade the criterion, the lower-level principle, the higher-level principle, and the value a "D". As an approach to answering this question, we focus on our case studies.

3.2 Aggregating and the role of the second case study

For the sake of illustration of a realistic scenario, the example is on the principle of non-discrimination (under the value of justice). We test this example with our two case studies. The second case study can be referred to an entertaining context with the aim of swapping faces in pictures. The second case study, is of particular importance in that it allows not only for the application and testing of the VPCIO model with a single case study, but also for the assessment of the model's broader applicability in the context of another AI system. As already mentioned, the thesis in this paper is that even if we have a blueprint for transferring values and

[22] It is possible, and perhaps even likely, that in other AI contexts, the indicators related to the value of well-being will not need negative anchor indicators, since our interpretation of well-being goes beyond aspects of justice because it requires more than what is derived from aspects of justice. However, since a malfunction or an error in the context of this AI-assisted robot can affect a person's life, we think it is reasonable to think of higher standards of well-being in this case.

Table 2 The aggregation of well-being in the case of R 1.6 as a negative anchor indicator

Observable	Indicator	Criterion	Lower-level Principle	Higher-level Principle	Value
B C	EA 1.1 EA 1.2	EA 1 C	Enhancing Autonomy C	Beneficence*[1] B	Well-being B or D
B B A	EA 2.1 EA 2.2 EA 2.3	EA 2 B			
C	EA 2.1	EA 3 C			
A B A A A D A A	R 1.1 R 1.2 R 1.3 R 1.4 R 1.5 R 1.6 R 1.7 R 1.8	R 1 D	Resilience / Reliability / Robustness C or D	Non-maleficence / Prevention of harm *[2] B or D	
A B	R 2.1 R 2.2	R 2 B			
A	DU/MA 1.1	DU/MA 1 A	Dual Use / Misuse Awareness A		
A	EF 1.1	EF 1 A	Ecological Flourishing A	Beneficence + Non-maleficence / Prevention of harm	
A B	EF 2.1 EF 2.2	EF 2 B			
A	EF 3.1	EF 3 A			

*1 The aggregation of 'beneficence' is based on that of 'enhancing autonomy' and 'ecological flourishing'. *2 The aggregation of 'non-maleficence/prevention of harm' is based on that of 'resilience/reliability/robustness', 'dual use/misuse awareness' and 'ecological flourishing'

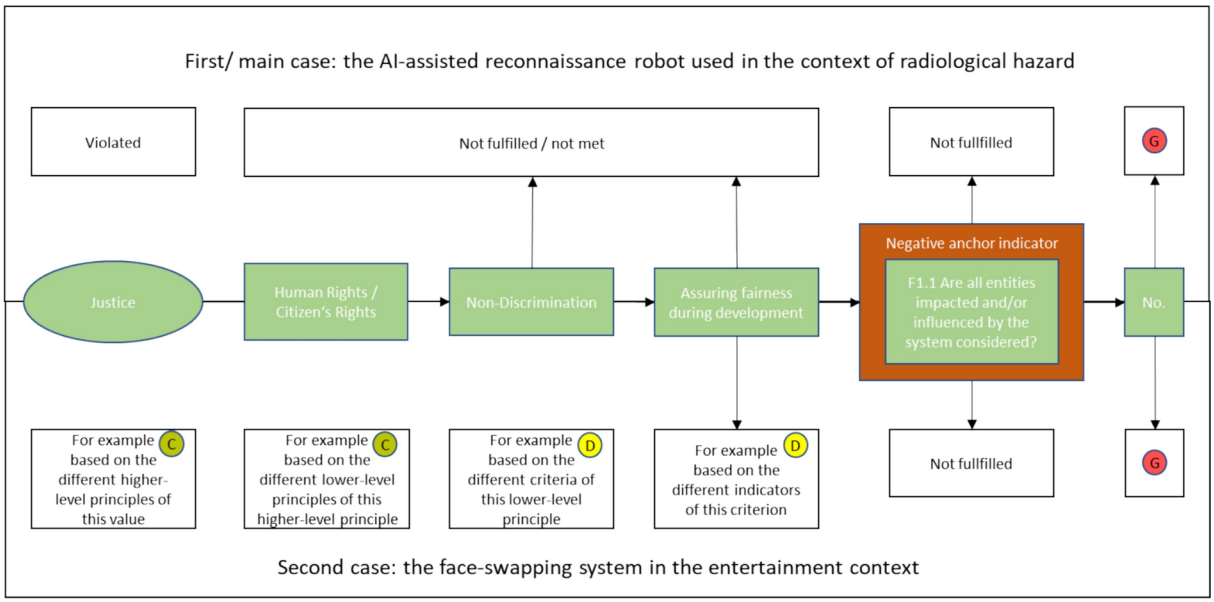


Fig. 7 The relevance of a negative anchor indicator

principles, we need to adapt and perhaps modify it to specific contexts and specific AI use cases.

It may be the case that the system presents a more favorable outcome when the individuals in question possess a similar skin tone. In other words, the result is of a lesser quality if one face is that of a person of color and the other face is a white person. In this scenario, the criterion 'Assuring fairness during development'[23] [9, p. 30] is not met due to the failure to fulfill a negative anchor indicator. If we posit that this non-fulfillment should be extended to the other level, the value of justice is not met. Such findings could result in a recommendation against purchasing or utilizing the face-swapping AI product, or even a complete ban on its market availability. Conversely, if the argument is made that the non-fulfillment should not be extended, the rate of non-fulfillment of the criterion is aggregated with the other evaluated criteria. This is the case in Hallensleben et al. [10, 9]. Some might suggest that in this scenario it would be reasonable to claim that the rate does not need to be extended to the other levels. This is due to the fact that the system in question has already been developed and is utilized for the purpose of private entertainment. Furthermore, this function can be achieved without significant negative effects on society,[24] even if the results vary in quality. In this case, however, it should be made transparent that the rate of the principle of non-disrcimination is low.

With respect to the main case study, which pertains to an AI-assisted reconnaissance robot in a potential radiological threat scenario, a system is utilized for the detection of a range of objectives, including danger symbols and human beings.[25] Consider the following fictional scenario: It can be reasonably assumed that the system would not detect people of color with the same reliability as other people. In fact, the probability of detecting people of color is 50%, while the probability of detecting other people is 95%. Accordingly, the criterion 'Assuring fairness during development' is not met due to the failure to fulfill the negative anchor indicator. It seems reasonable to suggest that the rate should be expanded to the other levels as the system is developed to detect people in scenarios where it may be a matter of life and death. In this scenario, it can be argued that it is ethically unacceptable to use a system that does not act in accordance with the principle of non-discrimination. This claim can only be supported in the aggregation of the VPCIO model if the non-fulfillment of the criterion is extended to the other levels (Fig. 7).

In this diagram, only those parts of the VPCIO are illustrated that are relevant for the assessment of the criterion'Assuring fairness during development' in the case described.

These examples for the unfulfilled criterion 'Assuring fairness during development' based on a negative anchor indicator show that the ethical evaluation of the aggregation type can vary depending on the context.

[23] We adopted the indicators for the criterion 'Assuring fairness during development', and the criterion itself, from Hallensleben et al. [9].

[24] This does not mean that the face-swapping system can never have a big impact. Imagine that the system is misused to swap faces in photos without consent, to distribute these manipulated photos without making the manipulation (with an AI system) transparent, this can have serious negative consequences and should be avoided.

[25] The case study is closely related tothe KIARA project (footnote 7), but the findings, etc. are thought experiments, and as such fictional.

Similarly, discrepancies in measurement may emerge within the same AI system, but across different stakeholder groups. This can influence a negative anchor indicator and a skippable indicator within the principle of transparency with respect to different groups of stakeholders, for example, operators, users, and affected individuals in the public.

3.3 The role of different stakeholders

Let us think again about the different stakeholder groups. In the main case study of an AI-assisted robot, for example, there are directly and indirectly affected civilians and the operator. If we think about the principle of transparency, different parties may need different levels of transparency to understand the information relevant to them. In the second case study of a face-swapping system, for example, there are the user and the people who are photographed and whose photos are used for face-swapping. When we think about the principle of privacy, different parties may need different levels of privacy. With the VPCIO model and the AI ethics label, it is possible to address these different needs by redefining negative anchor indicators in a specific use case of an AI system. The redefinition makes it possible to consider which observable (perhaps a "C" rating) is necessary and to what level (criterion or even value) the score should be expanded if it is not met. These points are also relevant for the positive anchor indicators and the skippable indicators.

However, when we consider a broader notion of stakeholders, we also think of the developer. The developer needs to know how to address ethical aspects. To understand this, the VPCIO model is helpful, as it illustrates the path from abstract requirements to concrete requirements (observables) that can be considered in design and development. In order to apply the VPCIO and the AI ethics label to a specific case, it is not useful to have different aggregations of the model for each stakeholder group. Therefore, it is important to either aggregate only the lowest rate[26] or to define the observables specifically for the different stakeholder groups[27].

Another stakeholder in a broader sense is the regulatory and legal system. To be more precise, these stakeholders are those who set the rules and law. First, the legal system needs to define rules that include ethical considerations to ensure AI ethics. Second, and analogous to the developer, the legal system needs to know how to comply with ethical aspects. The VPCIO model is adequate for this, as it allows a legally appointed oversight body to assess whether the use of an AI system follows the rules. A rule may be "The use of an AI system should be transparent", which is a possible formulation of the principle of transparency embedded in the VPCIO and therefore allows to assess (with the AI ethics label) the fulfillment of this principle formulated in the mentioned possible legal rule.

The findings indicate that the VPCIO model requires adaptation and that it is not feasible to utilize the same model for all AI systems and potential applications. It is necessary to reconsider the three different features, namely the three special indicators, for each AI system intended for a specific use. It is similarly crucial to determine the minimum acceptable scores and whether they should be extended to higher levels or aggregated in the manner described by Hallensleben et al. [10, 18].

It is important to note that the application of the VPCIO model to an AI system is not solely for internal evaluation. Rather, it is already intended by Hallensleben et al. as a tool to enable users of the system to gain insight into AI ethics compliance [10]. With the help of this tool, namely the VPCIO model, an AI ethics label can be obtained that provides the user with exactly this insight. In the absence of such insight "it remains crucial to communicate an AI system's ethical characteristics in a way that citizens, users, and consumers can easily understand. The same applies to policymakers, regulators or standard-setting bodies." [ibid., p. 31]. The idea of Hallensleben et al. is to provide an AI ethics label that offers an intelligible ethical summary of the respective AI system. This is intended to be "one among many possibilities how our VCIO model can be used for AI regulation" [ibid.]. Our modification of the VPCIO model towards an ethically measurable model, along with the refinements of related aspects and features based on the consideration of two case studies, aims to enhance the understanding of ethical aspects and their relations. This is to be achieved not only for the benefit of the developer but also to provide a refined version that can be used as a standard for AI regulation.

4 Conclusion

Given the potential of AI technologies to negatively impact people, there is a clear need for AI ethics, particularly the implementation of ethical aspects in AI systems. This paper addresses the question of how to transfer values and principles to an AI system in a way that is understandable and evaluable for different stakeholders, e.g. users or an oversight

[26] In the transparency example, the operator needs a higher level because it is most important for him to understand the system decisions. This would mean that if we determine that an observable should be rated "C" to ensure that the operator can deal with it, that is the rating that should be aggregated and not the "D" rating that would be sufficient for an affected person.

[27] In the example of transparency, there might be an observable such as "Are the processes and results of the AI system transparent to the operator during operation?" and another such as "Is it transparent to an affected person that an AI system is being used and for what?".

body. Ethical recommendations need to be understandable to ensure that both AI system developers and legislators can engage with them. The abstract nature of ethics, in particular ethical values and principles, poses challenges in their direct applicability to developers and their incorporation into legislation. The research background for this paper was already identified, shared and systematized ethical values and principles [3] and the so-called VCIO model [10], which provides a path from the abstract to the concrete by providing values, criteria, indicators and observables. Starting from Hallensleben et al.'s model, we offer a modified model that is consistent with the shared ethical aspects of the AI ethics guidelines (see above). Our modification includes these shared aspects, namely ethical values, which we interpret as desirable states, and ethical principles, which we interpret as action-guiding. By extending the model to include the category of principles, they can also be understood as links between values and criteria, so we present a VPCIO model.

In approaching the implementation of these ethical aspects, our approach was developed with the first case study, namely an AI-assisted robot used for reconnaissance and defense against acute radiological hazards. In this sense, we have adopted shared ethical aspects in the VPCIO model.

Based on the three values we have identified, namely 'well-being', 'justice' and 'understanding', we have developed novel criteria for the identified principles, with a particular focus on 'well-being' and 'understanding', as these values differ from those presented by Hallensleben et al. Furthermore, illustrative examples of indicators and observables have been provided.

With this VPCIO model, we provide an ethical framework for the use of AI systems that is transferable to practical requirements. The objective of our main case study was to identify specific requirements. One result of this focus is that the level of a required principle may vary with different stakeholder groups, which can be illustrated at the level of observables, as they are distinguished into seven levels. At this point, the second case study, a face-swapping system, is additionally considered in order to test the completeness of values, principles, criteria, indicators and observables, but also to evaluate whether the required level of their fulfillment varies with different contexts. The results of the comparison of the two case studies confirm the thesis that we need to adapt and possibly modify the indicator system (VPCIO) to specific contexts and specific AI use cases.

Nevertheless, the explanations of the model do not yet show how the fulfillment of the requirements can be assessed to ensure AI ethics. Therefore, we recommend a specialized and modified aggregation based on an aggregation method, namely the AI ethics label, presented for the VCIO model [9].

In the aggregation process, the observables are translated into letters from A to G, which are then aggregated to the different aspects, namely from indicators to criteria and from there to values. Our consideration of the two case studies showed that the adaptation of aggregation features is context dependent. Gaining such insights into AI ethics compliance, this insight can be realized with the AI ethics label [10, 9]. We argue for redefining and adapting elements of the AI ethics label that are specific to different AI systems in different contexts with respect to different stakeholders. It provides an understandable ethical summary that can be used for AI regulation. It therefore has legal significance, which can ensure AI ethics by appointing an oversight body that assesses the fulfillment of ethics with the aggregation of the VPCIO model for an AI system in a specific use case. It is also relevant to the developer by providing design indicators that are particularly relevant to developers and by helping to understand how to comply with abstract ethical aspects. The aggregation of the VPCIO in a specific case is also useful for the user, as it illustrates the system's accordance with ethical aspects. By outlining the minimum conditions for the acceptability of AI systems and proposing an aggregation methodology for an AI ethics label, we have developed a method for summarizing and aggregating the fulfillment of ethical aspects that can be useful for different stakeholders such as users, developers and the legal system or an oversight body. The ethical consideration is not complete with that, because the ethical acceptability varies with different contexts involving different stakeholders. Therefore, we need to address ethical requirements specific to certain contexts and stakeholders.

Funding Open Access funding enabled and organized by Projekt DEAL.
KIARA project: Federal Ministry of Education and Research (BMBF), 13N16277.

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

Declarations

Competing interests The authors declare no conflict of interest.

References

1. Asilomar conference (Beneficial AI), 'Asilomar AI Principles', Future of Life Institute. https://futureoflife.org/ai-principles/. Accessed: 08 Oct 2021 (2017)
2. Abrassart, C. et al.: Montréal declaration for a responsible development of artificial intelligence. Announced at the conclusion of the Forum on the Socially Responsible Development of AI. https://www.montrealdeclaration-responsibleai.com/the-declaration (2018)
3. Martin, D., Hillerbrand, R., Schmidt, M.W. (forthcoming): Comparing AI Ethics and AI Regulation: Ethical Values and Principles and the Case of Well-being. In: Müller, Vincent C.; Dung, Leonard; Löhr, Guido and Rumana, Aliya (eds.) (forthcoming), Philosophy of Artificial Intelligence: The State of the Art (Synthese Library, ed. Otávio Bueno, Berlin: SpringerNature)
4. Beauchamp, T.L., Childress, J.F.: Principles of biomedical ethics, 8th edn. Oxford University Press, New York (2019)
5. Braun, M., Hachmann, C., Haack, J.: Blackouts, restoration, and islanding: a system resilience perspective. IEEE Power Energy Mag. **18**(4), 54–63 (2020). https://doi.org/10.1109/MPE.2020.2986659
6. Floridi, L., et al.: AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Mind. Mach. **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
7. Flynn, J.: Theory and bioethics. In: The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2022/entriesheory-bioethics/ (2022) Accessed: Jun. 03, 2024
8. Greco, J., Pinto de Sa, L.: Epistemic value. In: Routledge Encyclopedia of Philosophy, 1st ed., Routledge, London. https://doi.org/10.4324/0123456789-P073-1 (2016).
9. Hallensleben, S., Hauschke, A., Hildebrandt, S.: VCIO based description of systems for AI trustworthiness characterisation, VDE Verband der Elektrotechnik, Offenbach am Main, VDE SPEC VDE SPEC 90012 V1.0 (en). https://www.vde.com/resource/blob/2176686/a24b13db01773747e6b7bba4ce20ea60/vde-spec-vcio-based-description-of-systems-for-ai-trustworthiness-characterisation-data.pdf (2022) Accessed: Mar. 16, 2023
10. Hallensleben, S. et al.: From principles to practice—an interdisciplinary framework to operationalise AI ethics. VDE, Bertelsmann Stiftung, Frankfurt a. M./Gütersloh. https://www.ai-ethics-impact.org (2020)
11. High-Level Expert Group on Artificial Intelligence set up by the European Commission, 'Ethics guidelines for trustworthy AI', European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai(2019)
12. House, T. W.: Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/ (2023) Accessed: Nov. 22, 2023
13. IEEE, 'Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, version 2', Version 2. [Online]. Available: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf (2017)
14. IEEE: Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, version 1. Springer, Cham (2019) https://doi.org/10.1007/978-3-030-12524-0_2.
15. IEEE: IEEE Standard for Transparency of Autonomous Systems (2021) [Online]. Available: https://standards.ieee.org/ieee/7001/6929/
16. IEEE: IEEE Position Statement. Artificial Intelligence. IEEE, 18029 (2019)
17. IEEE: AI ethics in the public, private, and NGO sectors: a review of a global document collection. In: IEEE Transactions on Technology and Society, 1st ed., vol. 2 (2021)
18. Kant, I. Critique of practical reason, Revised edition. In: Cambridge texts in the history of philosophy. Cambridge University Press, Cambridge (2015)
19. Korsgaard, C.M.: Two distinctions in goodness. Philos. Rev. **92**(2), 169 (1983). https://doi.org/10.2307/2184924
20. Larmore, C.: Right and good. In: Routledge Encyclopedia of Philosophy, 1st ed. Routledge, London (2016). https://doi.org/10.4324/9780415249126-L087-1.
21. Misselhorn, C.: Artificial moral agents: conceptual issues and ethical controversy. In: Voeneky, S., Kellmeyer, P., Mueller, O., Burgard, W. (Eds.) The Cambridge Handbook of Responsible Artificial Intelligence, 1st ed. Cambridge University Press, 2022, pp. 31–49. https://doi.org/10.1017/9781009207898.005.
22. OECD: Recommendation of the Council on Artificial Intelligence (2019)
23. Regulation (EU) 2024/1689. http://data.europa.eu/eli/reg/2024/1689/oj/eng (2024) Accessed: Jul. 26, 2024.
24. Schroeder, M.: Value theory. In: The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/value-theory/(2021). Accessed: Aug. 11, 2022
25. UNESCO: Recommendation on the Ethics of Artificial Intelligence (2022)
26. United Nations Department of Public Information, NY, Universal Declaration of Human Rights (1948)
27. Véliz, C.: Moral zombies: why algorithms are not moral agents (2021)
28. van de Poel, I.: Values and design. In: The Routledge Handbook of the Philosophy of Engineering (2020)