AI and Ethics
https://doi.org/10.1007/s43681-023-00330-4

ORIGINAL RESEARCH

# Ethics by design for artificial intelligence

**Philip Brey**[1] · **Brandt Dainow**[2]

Received: 3 April 2023 / Accepted: 2 August 2023


**Abstract**
In this paper, we present an approach for the systematic and comprehensive inclusion of ethical considerations in the design and development process of artificial intelligence systems, called Ethics by Design for AI (EbD-AI). The approach is the result of a three-year long research effort, and has recently be adopted by the European Commission as part of its ethics review procedure for AI projects. We describe and explain the approach and its different components and its application to the development of AI software and systems. We also compare it to other approaches in AI ethics, and we consider limitations of the approach as well as potential criticisms.

**Keywords** Ethics by design · Design ethics · Design methodologies · Ethical guidance · Human agency · Fairness · Transparency · Privacy · Accountability · Well-being

## 1 Introduction

Ethics by Design is an approach for systematically and comprehensively including ethical considerations in the design and development process of new technological systems and devices. Although the approach can be applied to any technology, historically its focus has been on the design of AI systems. In this essay, we present and critically discuss a full-blown approach in Ethics by Design for AI.

The concept of Ethics by Design has its roots in European studies in computer science, ethics and responsible innovation of the mid and late 2010s [1, 2]. Possibly, it is a generalization of the notion of "privacy by design" [3]. The central idea behind it is that technology can be made ethical through its design process. The concept acquired a boost in Europe through a 2019 resolution of the European Parliament regarding the need for a comprehensive artificial intelligence policy, which stated that the Parliament "believes that any AI model deployed should have ethics by design" [4].

✉ Philip Brey
p.a.e.brey@utwente.nl

Brandt Dainow
bd@thinkmetrics.com

[1] University of Twente, Enschede, The Netherlands

[2] Computer Science, National University of Ireland Maynooth, Maynooth, Ireland

While the concept of Ethics by Design for AI was thus established by around 2020, both in research and in policy, our assessment is that no developed approach had been presented so far. Past publications presented criteria and proto-approaches, but not a full approach that could be used in actual design processes. The 2018 article by Dignum et al., "Ethics by design: Necessity or curse?" provides a definition of ethics by design and proposes ethical principles and issues for Ethics by Design, but then does not proceed to provide an approach for it, although it reviews some ideas from a workshop that could provide building blocks for an approach [1]. The 2018 article by d'Aquin et al., "Towards an 'Ethics by Design' Methodology for AI Research Projects," [2] comes closer to defining an approach to Ethics by Design. It proposes requirements for an Ethics by Design methodology and a set of guiding principles towards it. But as it acknowledges, these are just steps towards an Ethics by Design approach, and the paper does not present a full approach.

In the period 2019–2021, the task to develop a complete approach was taken up in a collaboration between two EU-funded projects with a focus on ethics of AI: the SHERPA and SIENNA projects. These were the two largest EU-funded projects at the time focusing on ethical and human rights aspects of AI. With the approval of the European Commission, they set out to develop an Ethics by Design approach for AI in 2019.

Published online: 21 September 2023

The first version of this approach was developed within the SHERPA project [5], and a subsequent, improved version in the SIENNA project [6]. This latter version also became the basis for a guidance document which was published as part of the ethics review process for AI projects under the Horizon Europe funding scheme. We bear responsibility for the final version of the SIENNA/SHERPA Ethics by Design approach, as we are the authors of the final report on Ethics by Design and also worked with the European Commission to develop the guidance document on ethics by design for European funding [7].

We call this approach "Ethics by Design for AI" or "EbD-AI" in short. In this paper, we present our approach and situate it within the broader landscape of approaches in AI ethics. We will also discuss its strengths as well as its limitations.

## 2 Ethics by design for AI and its application

EbD-AI is an approach to AI ethics which is based on the conviction that technology is not neutral and that values can, to some extent, be embedded into the design process. A key premise here is that design choices are not morally neutral but can have significant ethical consequences. Granted, many of the consequences of a new technology are partially or wholly dependent on how and in what context it is used. However, design matters as well: design choices can sometimes generate particular consequences across a broad range of uses or use contexts. For example, an app that is designed to automatically collect and disseminate personal information about its users, without informed consent, violates privacy in a way which is largely independent of how and when it is being used.

The idea that values can be embedded in design is already somewhat familiar within computer science [8]. The idea of Privacy by Design [9], in particular, is already well established, and the idea of Secure by Design has also recently been established [10]. More recently, Transparency by Design has been proposed as an approach [11]. Beyond computer science, Safety by Design (or Safe by Design) is an approach which incorporates safety aspects of a system throughout the design process [12, 13]. Design for Sustainability does the same for sustainability [14]. These approaches demonstrate that a wide range of values can be taken into consideration during design. The goal of these approaches is often to bring it about that these values are actually realized or upheld. A guarantee for this cannot be given, because the application and use of systems introduces new variables that can often override these efforts. It is, therefore, proper to say that incorporating consideration of an ethical value during the design process increases the likelihood that this value is realized in the resultant system. However, there will always be uses or use contexts in which the desired value is not manifested.

Ethics by Design for AI provides engineers with specific tangible tasks which must be accomplished during system development. However, it does not specify how those tasks should be accomplished, because this will depend on the application to be developed and the organization doing so. The premise of the Ethics by Design for AI approach is that it is more effective to provide engineers with pre-identified tasks to be performed, rather than ask them to undertake philosophical deliberations. Engineers generally do not possess skills of ethical analysis and deliberation to recognize, assess and mitigate ethically problematic AI, while they are also faced with challenges that are likely to make AI systems have ethical issues, such as a lack of diversity in their field and biases in input data.

It is, therefore, insufficient to ask AI engineers to improve the ethical status of AI systems without providing suitable tools. The solution we propose is to add a framework in the form of a set of tasks which can be appended to their development methodology. Engineers are familiar with methodologies and frameworks for specific classes of concern, such as reliability. EbD-AI moves ethical concerns to the same level as such concerns so that ethical issues become a routine part of system development. We should emphasize that this approach works well for routine ethical issues in design. For special ethical issues, a reflective and deliberative approach is preferred. We recommend that an ethicist is connected to AI development projects who can do an initial ethical impact or risk assessment for the project and identify and reflect on such ethical issues with the team.

In the EbD-AI approach, the focus is on instantiating core moral values necessary for a design to meet ethical standards, ranging from privacy to fairness. As with many ethical terms, their exact meaning within a specific application will vary, or even be debatable. It will often be for the organisation to develop their own ethical stance regarding what a particular value means with regard to the specific application under development. How this is done is not a concern of the EbD-AI approach, which merely lays down *what* must be done, not how to make it applicable to all AI development organisations.

Let us now turn to our proposed EbD-AI approach itself, which ensures ethical matters are addressed throughout the full length of the development process. Requirements will apply not only to the AI system itself, but also to some of the processes and the tools used in its development. First, foundational moral values and principles of a general nature are converted into ethical requirements for the specific AI system to be developed, then it is determined how to build the system in a manner which instantiates them. Thus, ethical requirements are translated into concrete tasks, goals, tools, functions, constraints, and the like Fig. 1.

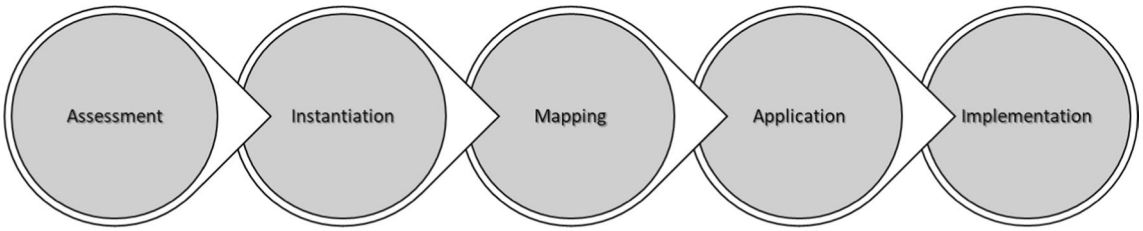


**Fig. 1** The steps on how to apply the Ethics by Design framework

We now offer an outline of how EbD-AI is used in AI system development. This is not a description of how to apply EbD-AI *during* the development process, which is discussed in Sect. 4 "Implementation" below. Instead, this section summarizes the main steps in the process of applying EbD-AI to any development. The EbD-AI approach is applied through five steps:

1. Assessment: First, the system's objectives are assessed against the foundational moral values (listed in Sect. 3 "Ethical Framework). If any core moral values are violated, the AI, as envisioned, is already unethical. If this is the case, the overall objectives of the application should be reconsidered.
2. Instantiation: Next, these moral values are instantiated as characteristics the AI system should possess. This provides ethical design requirements. How ethical design requirements are developed depends on the development methodology, organizational structure, and the nature of the AI system. It is not a part of the EbD-AI approach to determine how EbD-AI is introduced into any organization. Some tasks required by the EbD-AI will clearly demand appropriate organisational support. However, there will always be many ways to implement them.
3. Mapping: The third step of the Ethics by Design approach involves mapping these high-level ethical design requirements into specific procedures and actions to be taken during the design process. These will be implemented through various means, including system functionality, data structures and organizational measures. In some cases, ethical requirements may require additional functionality; in other cases they may act as constraints on what functionality is permissible. Many will call for additional organizational processes. For example, to avoid algorithmic bias, one should undertake a formal bias assessment of data before it is used.
4. Application: Because Ethics by Design is a high-level approach which can be mapped onto any development methodology, the fourth step is to determine where in one's own methodology each ethical requirement will be handled and how. The Ethics by Design approach facilitates this by offering a generic model of development into which values have already been translated into AI-specific requirements. The process of actually implementing EbD-AI within one's chosen methodology thus becomes one of mapping this generic model onto that methodology. Most organizations will need to reorganize their development processes to some degree. For example, version control systems may need modification to track ethical issues. Such change will work best if developers are encouraged to lead these changes [15].
5. Implementation: The final step is simply to implement Ethics by Design protocols during the development process in accord with the mapping done in the previous step. While this step can be described easily, it forms the bulk of the work when developing an AI system. As such, it is described in its own section below (see Sect. 4: "Implementation").

## 3 The ethical framework for ethics by design for AI

In this section, we present core moral values relevant to AI. These values form an "ethical framework," since they are the values and norms that inform the design of AI systems. As explained in the previous section, we start with high-level values for AI, which are then translated into design requirements for AI systems, which are further translated into specific measures to be undertaken at specific points in the design process, as described in Sect. 4 "Implementation".

The values in EbD-AI reflect an emerging global consensus regarding which moral values should govern AI. In recent years, numerous international organisations and national governments have proposed ethics guidelines for AI that specify these ethical values, including the Institute of Electrical and Electronics Engineers (IEEE), the European Commission's High-Level Expert Group on AI (AI-HLEG), the Organisation for Economic Co-operation and Development (OECD) and UNESCO. There is a strong agreement between these documents [16] that the following six values should be paramount in the development and use of AI: freedom, privacy, fairness, transparency, accountability, and

well-being (of individuals, society, and the environment).[1] Many of these values correspond to human rights affirmed in international and national laws and to those general ideas regarding avoiding harm and doing good, which are widely shared across cultures.[2]

The EbD-AI approach endorses these six values,[3] and, given that these guidelines were developed in a European context, we use the particular formulation of them by the EU-endorsed High-Level Expert Group on Artificial Intelligence (AI-HLEG). We now present a summary of the six values and the main design requirements that follow from them.[4]

### 3.1 Human agency

Human agency encompasses three values that are recognized as human rights: freedom, autonomy and dignity. Respect for human agency involves respecting for human being's right to have their own thoughts, make their own decisions and carry out their own actions.

**Design requirements**

1. An AI system should not be designed or used in a manner which deprives people of the ability to make decisions which they should be able to make for themselves.
2. An AI should not be designed or used in a way that results in the reduction of basic human freedoms, including freedom of movement, assembly, speech and information.
3. AI systems should not be designed or used to subordinate, coerce, deceive, manipulate, objectify or dehumanize people.
4. AI systems should not be designed or used to create addiction to the system or to the services which it provides.
5. AI applications should be designed to give system operators and (as much as possible), end-users the ability to control, direct and intervene in operations of the system.
6. AI systems should never make the final decision about important issues of a personal, moral or political nature. They may recommend, but the final decision must always be made by a human.

### 3.2 Privacy and data governance

This principle affirms the right to privacy of all human beings, and affirms the importance of data governance, including measures to support the quality and accuracy of data, access to data, and other data rights, such as ownership. Requirements apply to both the data used during the design of the AI system and data generated by it when in use.

**Design requirements**

1. The rights of data subjects should be respected during the processing of personal data.
2. To ensure accountability, where personal data is processed by an AI, there must be ways to demonstrate how it ensures lawfulness, fairness and transparency of that data processing.
3. Measures must be in place to safeguard the rights of data subjects through technical measures, such as anonymization, as well as through organisational measures, such as access control systems.
4. Whenever relevant, AI systems must support the right of an individual to withdraw consent for the use of their personal data.
5. Data should be acquired, stored and processed in a manner which can be audited by humans.

### 3.3 Fairness

This value implies that people should be given equal rights and opportunities and should not be advantaged or disadvantaged undeservedly. Fairness implies the absence of any form of discrimination, as well as support for diversity and inclusion.

**Design requirements**

1. AI systems should avoid algorithmic bias, including bias in input data, modelling and algorithm design.

[1] A seventh frequently mentioned value is (cyber)security. We do not include it in our list, however, because we believe it is already sufficiently accounted for in regular development processes.

[2] Good definitions of these values, and other terms relevant to Ethics by Design, can be found in the Assessment List for Trustworthy Artificial Intelligence (ALTAI) which has been produced by High-Level Expert Group on Artificial Intelligence [17].

[3] It is important to understand the purpose of this paper is not to argue in favour of these values. They have already been adopted as the foundational values of the Ethics by Design for AI approach by the EU. The aim of this paper is to report them and describe how to apply them.

[4] Ours is not the first attempt to operationalize high-level ethics principles for AI. What is special about it, though, is that it operationalizes these specifically for design and development of AI systems, and as design requirements that are further operationalized as specific design actions. Compare this with the AI-HLEG, which has operationalized its ethics principles into a checklist called the Assessment List for Trustworthy Artificial Intelligence (ALTAI) [17]. This checklist does not differentiate between requirements for design, deployment and use, and also does not differentiate between mid-level requirements and low-level procedures and actions, as we do.

2. AI systems should, to the extent relevant and possible, be universally accessible and offer the same functionality and benefits to end-users irrespective of their different abilities, beliefs, preferences or interests.
3. AI systems should, to the extent possible, be designed to avoid negative social impacts on social groups, especially protected social groups.

### 3.4 Individual, social and environmental well-being

According to this principle, AI systems should contribute to, and not harm, individual well-being, the quality and functioning of society, and the quality of the environment.

**Design requirements**

1. AI systems should be safe to use and should not have a propensity to harm or significantly reduce the health and physical or psychological well-being of any stakeholders (users, clients, data subjects, and other affected parties).
2. AI development should be mindful of the principles of environmental sustainability, both regarding the system itself and the supply chain to which it connects.
3. AI systems should not negatively impact the quality of communication, social interaction, information, social relations or democratic processes; for example, by amplifying fake news or segregating people into filter bubbles.

### 3.5 Transparency

This refers to the idea that the purpose, inputs and operations of AI applications should be knowable and understandable to its stakeholders. This is so they can understand how, and for what purpose, these systems function and how their decisions are arrived at.

**Design requirements**

1. It must be made clear to users that they are interacting with an AI system—especially for systems that simulate human communication, such as chatbots.
2. The purpose, capabilities, limitations, benefits and risks of the AI system and the decisions it makes must be openly communicated to all stakeholders.
3. AI systems must be constructed so that people can audit, query, dispute or seek to change its activities. This includes organizational processes by which the operators can receive and assess requests from third parties.
4. When building an AI system, one should consider what measures will enable the traceability of the AI system during its entire lifecycle, from initial design to post-deployment evaluation.
5. Whenever relevant, AI decisions should be explainable to users. Where possible this should include the reasons why the system made a particular decision. We recognize that this may not be possible with some systems. Nevertheless, the system (or those deploying it) should always have a mechanism by which to explain what the decision was and what data were used to make that decision. Explainability is especially important for systems that make decisions or perform actions for which accountability may be required, such as decisions that can cause harm or restrict an individual's rights.
6. AI development processes always involves making decisions about ethical issues, such as how to remove bias from a dataset. Transparency requires that development processes and tools record these ethical design decisions so that it is possible to understand how ethical obligations were met. This information may be required for audits, for disputing decisions made by the system or for correcting any ethical issues which arise after deployment.

### 3.6 Accountability and oversight

Accountability for AI applications means that actors involved in their development and operation take responsibility for the way that they function and for the resulting consequences. Human oversight means that humans are able to understand, supervise and control the design and operation of their systems. Oversight is a condition for accountability, since actors need it to have the information and influence that is needed for accountability.

**Design requirements**

1. AI systems should allow for human oversight regarding their decision cycles and operation, unless compelling reasons can be provided which explain why oversight is not required.
2. The deployment process of an AI system should include risk assessment. Procedures for mitigation after deployment should be in place from the moment the system is deployed.
3. AI systems should be auditable by independent third parties. The procedures and tools available under the XAI approach [18] support best practice in this regard. This is not limited to auditing the decisions of the system itself, but also the procedures and tools used during the development process. Where relevant and practical, the system should generate human accessible logs of the AI system's internal processes.

As a next step, each of these design requirements is translated into further procedures and actions at different stages of the design process, as explained in the next section.

## 4 Implementation of ethics requirements within specific development methodologies

To be useable in any design methodology, Ethics by Design offers a generic model of system development. This model has been intentionally designed so that it can be mapped onto any formal development methodology. The model frames system development in terms of six generic "phases." While they are, unavoidably, described in a linear sequence of discrete steps, they should be understood as classes of operation which can be mixed, rearranged and parsed as appropriate to fit into the development processes which one will actually use when building an AI system.

We present a summary of these phases below. We have summarized the tasks to be performed in each of these phases. These include both tasks directed at the implementation of the ethics requirements presented in Sect. 3 and tasks of a more general nature that promote ethical design. In our full approach, there are 63 tasks. While we will provide examples of specific tasks to be performed in each phase, space prevents us listing all these tasks in full. Detailed tasks lists for each phase can be found in the EU's official guideline document [7]. However, it is important to appreciate that we are not proposing a definitive task list which is applicable to all possible developments. Some tasks may not be relevant, while other tasks we have not listed may be required. Our aim has simply been to identify those tasks we consider essential, together with those which are relevant to the widest range of AI systems. This model complements that described in Sect. 2 "EbD-AI and its application". It contains the same processes, but describes them in more operational detail and pins them to specific steps during software development. Section 2 described the general approach to using EbD-AI. This section now describes one method for doing so, but others may be more appropriate for some cases.

To give an idea of what tasks look like and how they are related to ethical requirements, consider the following example: The principle of transparency contains an ethics requirement that it must be clear to users they are interacting with an AI system. This requirement is first considered during the specification of requirements phase, which requires that design specifications, constraints, selected resources and infrastructure must be assessed for compatibility with the ethics requirements. Thus, a requirement is added that the system have ways of informing users it is an AI system. This will require planning for specific components during the detailed design phase, possibly a notification system or a user acceptance form. It will also generate a second more general requirement that the system is not constructed in such a way that the interface or output could be mistaken for a human (as much as possible).

Next, there is a specific task during high-level design which instructs developers to ensure there is no aspect of the AI system which could be mistaken for a human. At the high-level design phase, designers will therefore need to evaluate all features and functions being contemplated for their potential to mislead people. Some may need modification, while others may require additional features to provide the necessary information to the user. In addition, functional components, such as user notification systems, need to be designed. These then need to be constructed during the development phase. Finally, during the testing and evaluation phase, there will be several related tasks, such as testing the functionality of user notification systems, but also general testing to establish whether users understand that they are interacting with a non-human agent and/or that a decision, content, advice or outcome is the result of an algorithmic decision in situations where not doing so would be deceptive, misleading, or harmful to the user.

Within each phase, we have determined specific tasks required to fulfil the high-level values and corresponding ethics requirements, listed above. Our following description of the phases in our generic model summarizes the key objectives of each phase and provides examples of some of the tasks which should be performed to meet those objectives Fig. 2.

It is an unavoidable consequence of the nature of "by design" approaches that many, if not all, of their requirements require interpretation. Requirements are not technical specifications. They are generic characteristics which relevant AI systems should exhibit. In the same way as requiring a software system should be reliable, requiring an AI system be ethically compliant does not specify exactly what functionality it should exhibit, how it should be constructed, or what organisational processes are required. AI is a general technology, and so any approach applicable to all AI must necessarily avoid domain-specific requirements, which would not be applicable to all. For example, the requirement that "AI systems should be safe to use and should not have a propensity to harm or significantly reduce health and physical or psychological well-being of any stakeholders" may legitimately be considered open-ended. This is intentional because we cannot anticipate every possible AI application which may one day be created. Thus the intention behind the requirements is that they be as applicable as possible by remaining at a general level.

These six generic "phases" are as follows:


**Fig. 2** The six "phases" of the Generic Development Model

1. Specification of objectives. This involves a top-level outlining of the purpose and desired capabilities of the system. It is important to ensure that these align with ethical standards. It is also important to consider any ethical issues which may occur during the development process, perhaps because of particular procedures or tools. The objectives of the system to be designed are evaluated against the six generic moral values listed above. Some may be more important than others or even have legal mandates. For example, the value of Data Governance includes compliance the GDPR, or its local equivalent. An example of a specific task to be performed to ensure compliance with the value of Data Governance is to assess whether plans for what data will be used are fair and appropriate. If the aim of the system is not compatible with these values, then it is not permissible to develop the system as envisioned. Another restriction concerns objectives that are likely to result in harm (physical, psychological, or financial) or damage (environmental or damage to social institutions). System objectives that would lead to discrimination or disadvantage for certain social groups should be modified to prevent this. In the case where a project is inherently incompatible with the ethical values, then the project should not proceed. However, it is also important to consider potential misuse of the system, whether intentional or accidental. In addition to assessing whether the system's objectives meet the ethical principles and requirements, the impact of a system on stakeholders permeates the moral values listed above. The inclusion of external stakeholders when setting both the system's requirements and the ethical principles thus becomes a critical ethical requirement from the earliest stages of design.
2. Specification of requirements occurs when the development methodology calls for the determination of technical and non-technical requirements for the system. This should include those requirements needed to achieve good ethical status. The aim of this phase is to produce a development plan, which includes design specifications, task deadlines and an assessment of the staff resources required. In most projects, organizations will already have a standardized set of tools. However, these may not sufficiently allow for an Ethics by Design for AI approach. This means it might be necessary to factor in acquisition or development of new tools to support EbD-AI. Moreover, additional requirements for human oversight and documentation are likely to mean that internal processes have to be changed. The requirement for auditability and accountability will most likely call for a formal, documented, assessment of the proposed design for compatibility with the ethical requirements. This should include an ethical risk assessment or ethical impact assessment, in which ethical risks are identified specific to the system to be designed. It may also require the production of an EbD-AI implementation plan which can ensure incorporation of the Ethics by Design for AI approach. It is advisable that planning at this stage includes an ethical compliance architecture which is embedded into the development process, especially if adopting EbD-AI for the first time. Since it may require specific tools or system functionalities, the development of an ethical governance model is appropriate at this time. Testing of the completed system will now require testing for ethical compliance, which will require new procedures and resources, so this should be planned for at this time.
3. The high-level design phase involves the creation of the system's high-level architecture, sometimes preceded by developing a conceptual model. The technical and non-technical requirements of the system, either of which can include ethical requirements, are incorporated into this phase. Key concerns at this stage include transparency, autonomy, privacy and fairness, which should be translated into specific functionalities within the system. It is the essence of EbD-AI that one ensure the ethical requirements are included in the list of system requirements just like any other, such as security or reliability. Ethical requirements should be treated exactly like traditional system requirements, meaning that they should be

programmed in the system organically, rather than being treated as special, ancillary or extraneous. Requirements such as auditability, traceability and accountability usually mean keeping records of internal data manipulation by the system. The requirement for transparency will demand that human oversight keeps an eye on these developments, while traceability and accountability require this oversight is documented. Consequently, a specific task required at this time is the design of mechanisms by which to document how data acquisition, storage and use happen. This needs to be auditable and must cover both the development process and use once operational. An additional task required in this regard is the design of ethical documentation systems which are sufficient to make ethical issues identifiable and their resolution traceable and explainable. The general concern in this regard is to ensure the system is transparent and that it supports the other ethical requirements, such as protection of personal data and ethical governance. In addition, explaining how the system design will promote universal accessibility is to be encouraged. This may include an accessibility assessment of the initial design. Where it is anticipated ethical requirements cannot be met, justification for this should be developed during this phase.

4. Data collection and preparation occurs when one's development methodology calls for the assembly and integration of data. The data collection and preparation phase is crucial, because incomplete or biased datasets are a major cause of unethical AI systems [19]. In the same way as newly written code is automatically presumed to contain bugs, it is essential to treat collected data as biased or incomplete until it has been tested and validated. Concrete measures must be taken to address this concern. The ethical requirement for AI systems to be explainable also means the measures taken to check and remediate data need to be documented and can be justified as sufficient and appropriate. Data may reflect the biases of a particular society, so there needs to be active proof that the data are representative or neutral before it can be used in the system. Care should also be taken to ensure that the learning or algorithmic manipulation processes do not introduce new biases. Ethical issues can also arise during the preparation of the data because even data cleaning can introduce problems [20]. As a result, steps should be taken to ensure that new biases are not introduced into the dataset or that other problems do not arise during use, such as de-anonymisation. It is, therefore, important to assess where ethical and/or data protection requirements could be violated, ensuring that the processing of data follows the GDPR or other relevant legislation. For example, if your planned system will process personal data, the values of data minimisation, traceability and accountability all require that your application is capable of demonstrating how you have incorporated the rights to data protection into your design and how it contains specific design features to enable this (such as the right of data subjects to have their data corrected or deleted). Similarly, transparency dictates you must carry out an analysis of the ethical risks related to the data processing and if needed, produce a risk mitigation plan.
5. The design and development phase occurs when actually constructing the system. Here ethical requirements become instantiated through the various construction processes, supported by appropriate tools and methods. Specific actions to incorporate ethical requirements are interpolated with other tasks. For example, transparency and accountability dictate you ensure the code is actively documented within the software program (as appropriate to the language(s) and methodology) and in appropriate ancillary documentation. It is important to ensure documentation is understandable to fellow programmers and accessible by them. It is highly likely that some procedures will need to be modified to build the ethical requirements into the system. In addition, it is likely some changes to organizational structure will be required to manage ethical requirements across the development lifecycle. For example, Google have developed an organizational framework they call SMATR (Scoping, Mapping, Artifact Collection, Testing and Reflection) to audit implementation of EbD-AI requirements throughout the development process, "all of which have their own set of documentation requirements and account for a different level of the analysis of a system" [21]. Supplementary tools will be required which programmatically support ethical values, such as the "Model Cards for Model Reporting" [22] and "Datasheets for Datasets" [23] toolsets. Additional documentation will be required both to ensure proper implementation of EbD-AI and to satisfy requirements such as accountability and traceability. Since the requirements of Ethics by Design for AI are widely recognized, formal structures for such documentation often exist, such as those of Explainability Fact Sheets [24]. It is important to note, as demonstrated above, that the requirements of Ethics by Design for AI are not new or innovative, but have been widely recognised for some years. As a result, no organization is required to invent everything from scratch. Tools, procedures, guidelines and organizational frameworks exist under various independent initiatives by which to satisfy these requirements. EbD-AI is simply an approach by which to organize and implement them. Most of these needs should have been recognized and planned for during the Specification of Requirements phase, if there is sufficient granularity in the plan-

ning. However, at least for the first project or two, it is likely that once development commences, things will not work as anticipated, or unforeseen needs will arise, such that project plans will need modification and additional tools, procedures or organizational structures will be needed. It is, therefore, important to add sufficient slack to project timescales for the organisation to learn how to do Ethics by Design for AI in its own unique context. Furthermore, as anyone who has moved a development team to a new methodology or coding platform will recognize, proper management of the human element is essential during this process [25, 26]. The ethical guidelines should be shared with all those working on the system, while managers should treat ethical requirements as fundamental to the system, not as questionable extras forced on them. In this respect, importing toolsets which can programmatically handle many of the requirements serves the dual purpose of shifting the burden of labour away from coders while also promoting an understanding that ethical considerations are now merely a standard part of the AI development universe.

6. The testing and evaluation phase involves verifying that the system meets the original objectives and requirements, which were determined in the specification phases (1 and 2 above). An ethical examination should be conducted during testing to verify the system complies with all its ethical requirements. If properly (and fully) implemented, the EbD-AI approach will have anticipated and addressed most, if not all, ethical issues before reaching this phase. It is important that real-world use of the system be assessed at this point. For example, since many ethical problems with AI involve unequal treatment of minority subsets of users [27], it is advisable to involve a good cross-section of potential stakeholders during testing. It might be the case that the functional requirements have been achieved, but not all the ethical requirements. For example, one task specified under the value of transparency is to test whether users fail to understand that they are interacting with a non-human agent. The aim of EbD-AI is to prevent such an outcome, so ethical issues should have been dealt with during the development process and ought not come up at this stage. A project ethical requirements checklist can be used to confirm the system's ethical compliance. It cannot be assumed that existing testing methods can assess ethical requirements adequately. As result, it is important during the Specification of Requirements phase to assess, and probably modify, the plans for final system testing. For example, if running beta testing with end users, reporting mechanisms need to contain provision for identifying and reporting ethical concerns. If it is determined that an ethical requirement has not been met, then this should be treated as one would treat any other type of bug and dealt with accordingly. It is essential that stakeholders are involved in this step because impact on stakeholders is a major concern of ethical AI. The understanding of the system's behaviour by its end-users should also be included in the testing process because ethical problems can arise when users misperceive what the system offers in some manner, most often by over-estimating its capabilities [28]. Certain key tasks are necessary to assess whether ethical requirements have been fully implemented: testing whether users understand that they are interacting with an AI, ensuring that human-useable auditability is built into the system, confirming that stakeholders and end-users understand the system's capabilities and limitations, ensuring there are processes by which end-users can report risks or biases in the system and formally attempting to predict the consequences of the system's functionalities. It is important that assuring ethical compliance is not left until the final testing phase. The whole point of EbD-AI is to ensure ethical compliance throughout the entire process. Many aspects of the system will be unchangeable by the testing phase. For example, oversight, transparency and human governance all require extensive logging in human-readable format by many elements of an AI system, from initial learning to decisions. It would be extremely difficult to add such functionality after a system is completed. It is, therefore, essential that compliance with the demands of EbD-AI is monitored and enforced throughout the entire development process. As development continues, deviations from the ethical requirements should be spotted and addressed as soon as possible. Doing so ensures minimal disruption to the development process. If done correctly, the final tests for ethical compliance should be mere formalities to confirm what has been monitored and built into the system from the beginning. EbD-AI is not something which can be isolated from the rest of the developers, but requires a cultural adjustment so that it permeates the entire development organisation.

The EbD-AI approach can be integrated into any design methodology from this generic model because the construction of any AI system must undertake the tasks listed above in some manner. Accordingly, these tasks need not follow a strictly linear progression. For example, in an AGILE environment, some phases would be iterated multiple times. Irrespective of the methodology being used, the objective of Ethics by Design for AI is to take ethical issues into account throughout the entire development process, treating them just like any other "more technical" requirements, such as reliability, usability or security.

The steps by which to integrate the Ethics by Design approach are as follows:

1. Map the tasks described in the generic model into one's chosen methodology.
2. Determine which ethical requirements are relevant within each stage of that methodology.
3. Use this to create a list of ethical requirements for each stage of your methodology.
4. Review the aim of the project, including data sources, functionality, output and deployment context. Determine if the list of ethical requirements is sufficient to cover all these aspects or if additional requirements are needed. This is particularly important if the system could have a significant impact on people's lives.
5. Implement formal systems to include EbD-AI into every element of your development methodology. Checklists of ethical requirements for each element are often a useful starting point. The first time any organization uses EbD-AI, it is likely to discover that additional tools and procedures are required. Fortunately, as illustrated above, the ethical requirements for AI outlined here are not unique to Ethics by Design for AI, but are widely recognized in the AI development community. As a result, many initiatives, such as XAI (eXplainable AI) [29], exist offering tools which can be used to incorporate EbD-AI tasks into any existing development methodology.

## 5 Strengths, limitations and the role of EbD in AI ethics

Ethics by Design for AI is an approach to AI ethics which aims to include ethical considerations in the design and development of AI systems in both a systematic and a comprehensive manner. We would like to contrast and compare our approach to several other approaches in AI ethics. First, our approach is different from efforts to develop high-level ethics guidelines for AI, such as the mentioned guidelines from organizations like UNESCO, IEEE, OECD and AI-HLEG. Our effort is, instead, to operationalize ethics guidelines to guide a specific practice, namely the design of AI systems.

Our approach is also different from approaches to ethical assessment of AI, which focus on identifying ethical issues in AI, but not necessarily on mitigation during the design process [30, 31]. It should also be distinguished from approaches that focus on the inclusion of ethical considerations in the deployment and use of AI systems. These are processes that are largely separate from systems design. We actually developed a separate approach for the ethical deployment and use of AI systems that is based on the same high-level principles as our Ethics by Design approach, but that are operationalized for the deployment and use of AI systems [32].

Let us now turn to approaches other than our own that also focus on ethical design and development of AI systems. One alternative approach is a research ethics approach for AI. This is any approach that focuses on ethics assessment of AI systems development project by a research ethics committee. In the review procedure, the committee reviews a project plan prior to the project's inception and assesses whether ethical issues can be identified or foreseen at this point and whether these are adequately accounted for in the plan. This process is quite different from Ethics by Design for AI, which is a continuous procedure rather than a one-time intervention, and which does not involve a research ethics committee in the process, but puts the largest responsibility for consideration of ethical criteria on developers rather than external parties.

We actually co-developed a research ethics framework for AI with the European Commission, and this approach is now part of its Horizon Europe ethics review framework [33, Annex 1, 34]. This is a framework that does not only cover AI systems development, but also fundamental research in AI. Under the Horizon Europe ethics review procedure, compliance to AI research ethics standards is mandatory for AI projects, and the use of the EbD-AI approach is recommended but not required. The research ethics framework for AI is based on the high-level principles used that are also used in the Ethics by Design for AI approach, consisting of a checklist with nine questions that inquire about specific ethical issues that may be at play. If any question is answered affirmatively, the researcher must detail how the issue will be dealt with in the project. We consider that Ethics by Design for AI and research ethics can have complementary roles for AI systems design. Research ethics allows for an initial screening for ethical issues, including ones that are not generic but specific to the proposed design project. Ethics by Design for AI subsequently takes these specific issues on board and ensures a systematic accounting for them and others over the course of the design process.

Ethics by Design for AI can also be contrasted with embedded ethics approaches for AI development, which focus on the inclusion of ethicists in design teams so that they can bring their expertise in ethics to bear on ethical issues and dilemmas in the design process [35]. The Ethics by Design for AI approach certainly does not preclude the inclusion of ethicists in design processes, and we in fact recommend that ethicists are involved at least in an initial ethical assessment, if not throughout the whole design process. Nevertheless, we hold that the inclusion of ethicists in design teams without the simultaneous inclusion of an EbD-AI approach will not lead to optimal results, since most ethicists do not have expertise in AI systems development and may therefore not be able to provide detailed recommendations to developers on what actions to take. The Ethics by Design for AI approach has already made many

translations of ethical criteria to specific design processes and actions and thereby bridges a gap between technical and ethical expertise which computer scientists and ethicist may otherwise find difficult to bridge.

Our approach can also be contrasted with approaches that focus on design for particular values. We already mentioned Privacy by Design, Secure by Design and Transparency by Design, amongst others. These approaches could, however, enrich Ethics by Design for AI by improving its methods for designing for particular values. Our approach is also different from approaches that focus on ethical design at the algorithmic level, such as ethical algorithm design [36] and algorithmic fairness [37], which can also further enrich and improve Ethics by Design for AI. It can moreover be contrasted to approaches that do not focus on the design of AI systems specifically, but of information systems more generally, such as value-sensitive design [38] and the IEEE Standard Model Process for Addressing Ethical Concerns during System Design [39]. Our approach is broadly compatible with the latter approach in particular, but differs due to its focus on ethical issues and design processes that are specific to AI.

### 5.1 Strengths, limitations and criticisms

We believe that an Ethics by Design approach provides the best approach for the incorporation of ethical criteria in the development of AI systems. It is, moreover, an approach that does not preclude other approaches that have been proposed for incorporating ethics, including research ethics, embedded ethics and approaches that focus on design for particular values or on ethical design of algorithms. In fact, we hold that combining these approaches with an Ethics by Design for AI approach can lead to better outcomes.

Let us now discuss some strengths and limitations of the Ethics by Design for AI approach that we propose, beginning with strengths. A first strong point is its intended integration with regular design methods. Our belief is that system developers are the individuals who ultimately make the decisions and carry out the actions needed to incorporate ethical criteria in design, and that they are more likely to do so if these criteria are integrated into their regular design methodology. This also has the advantage that development processes are not dependent on the presence of ethicists for ethical considerations to be included.

A related strong point is its operational detail. Our approach operationalizes ethical values and principles, not just by translating them into design requirements, but also by further translating them into detailed processes and actions to be taken at different phases in the design process. A final strong point is that the approach is flexible because it can be combined with a variety of existing design methodologies for AI systems. It also allows for combination with other approaches to ethical design, such as the ones mentioned previously.

Let us now turn to potential criticisms of our approach and resulting weaknesses. One criticism that we have heard made to our approach is that it preselects the values that should govern the development of AI systems and is, therefore, not flexible with respect to what values should be included in the approach. This is true, but we did not select the six values that we propose because we personally like them, but because there is a broad international consensus about their applicability to AI, as discussed earlier in the paper. Moreover, many of these values reflect basic human rights and principles of ethics about which there is a longer history of international consensus, as reflected, amongst others in the UN Declaration of Human Rights. Note that also our approach does not preclude the addition or subtraction of values. It is merely the case that we have made translations into ethics requirements, designs processes and actions only for the values that we have offered here. Indeed, it is the case that we have focused on these particular values precisely because of their widespread acceptance and therefore general consensus.

A second criticism is that the approach is too focused on the application of ethical principles and does not leave enough room for discovery of new ethical issues and for reflection and deliberation. We believe that there is plenty of room for these processes. However, we distinguish between "routine" and "special" ethical issues in design. "Routine" issues are issues that are likely to show up in any design project, such as generic privacy issues in relation to personal data or the possibility of algorithmic bias. We believe that such "routine" issues are best dealt with by making them part of normal design methodology and that the amount of reflection needed for them is limited.

In addition, our approach involves an ethical impact assessment for the specific design project early on in the process, which aims to uncover special ethical issues and is preferably carried out by an ethicist in collaboration with developers. These special issues may not be covered by the principles and requirements in place and may impose additional design requirements or specific actions required for mitigation. In addition, many of the actions that we recommend at different stages of the design process involve reflection before action is taken, such as processes that involve stakeholder consultation, which is a deliberative and reflective process.

A third potential criticism is that the approach is too complicated and cumbersome for designers. They could have to read a 30 page document (the length of our document included in the Horizon Ethics AI guidelines), map our generic design phases onto their own favored design methodology and then apply dozens of actions to implement the approach. Our reply is, first of all, that serious consideration

of ethical issues in systems development cannot be achieved by reading one-page documents and spending a few hours discussing ethics. A development project can take several person-years if all time is added up, and if even one half of a percent of this time is devoted to ethical considerations, that could still amount to weeks, if not months, of time. Moreover, learning the Ethics by Design for AI approach is a one-time process, as the skills are transferable to other projects.

Ideally, though, we think that EbD-AI should be taught as part of professional education, in accredited AI and computer science programs. It will be a long process before this is realized, but it is encouraging that there is growing recognition in the AI community that there should be design methodologies containing many of the principles we have proposed, including the already existing privacy by design methodologies and methodologies for algorithmic fairness, transparency and accountability. It is possible that in the future, these approaches will be incorporated into a more general Ethics by Design for AI methodology, but they may also remain as separate methodologies. In either case, our position is that they have a place in the curriculum of AI and computer science programs.

## 6 Conclusion

In this article, we presented an Ethics by Design approach for the development of AI systems, EbD-AI. The approach specifies recommended processes and actions at different phases in AI development, intended to incorporate six ethical values or principles: agency, privacy, fairness, well-being, transparency and accountability. These values are translated into design requirements, which are then translated into specific actions to be taken at different design phases. The approach always includes an ethical assessment of the specific system that is being designed and is intended to accommodate special, as well as standard, ethical issues in design. We also discussed how the approach relates to other approaches in AI ethics, and we discussed its strengths and potential weaknesses. The argument we have strived to establish is that Ethics by Design for AI is an indispensable approach for the incorporation of ethical considerations in the design process of AI systems and therefore should be seen as a central approach in AI ethics overall.

**Funding** This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreements No. 786641 (SHERPA) and No. 741716 (SIENNA). We wish to thank Tynke Schepers for editorial support.

## Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no conflict of interest.



## Reference lists

1. Dignum, V. et al.: "Ethics by design: Necessity or curse?," in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society 60–66 (2018)
2. d'Aquin, M., Troullinou, P., O'Connor, N. E., Cullen, A., Faller, G., Holden, L.: "Towards an 'Ethics by Design' Methodology for AI Research Projects," in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA. pp. 54–59. (2018) https://doi/.org/https://doi.org/10.1145/3278721.3278765
3. Cavoukian, A.: Privacy by design: The 7 foundational principles. Information and privacy commissioner of Ontario, Canada **5**, 12 (2009)
4. European Parliament, "A comprehensive European industrial policy on artificial intelligence and robotics," Feb. 15, 2019. https://www.europarl.europa.eu/doceo/document/TA-8-2019-0081_EN.html (Accessed Feb 15, 2023)
5. "Project Sherpa – Shaping the Ethical Dimensions of Smart Information Systems a European Perspective." https://www.project-sherpa.eu/ (Accessed Feb 22, 2023)
6. Brey, P., Dainow, B.: Ethics by design and Ethics of use in AI and robotics. In: Resseguier, A., Brey, P., Dainow, B., Drozdzewska, A., Santiago, N., Wright D. (eds.) Annex 2 of SIENNA D5.4: Multi-stakeholder Strategy and Practical Tools for Ethical AI and Robotics, pp. 33–73 (2021). https://doi.org/10.5281/zenodo.5536176
7. European Commission, "Ethics By Design and Ethics of Use Approaches for Artificial Intelligence (1.0)," European Commission - DG Research and Innovation, 22 2021. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf
8. Nurock, V., Chatila, R., Parizeau, M.-H.: "What does 'Ethical by design' Mean?," in Reflections on Artificial Intelligence for Humanity, Braunschweig, B., Ghallab, M. Eds. Cham: Springer International Publishing, 2021, pp. 171–190. https://doi.org/10.1007/978-3-030-69128-8_11.
9. Gürses, S., Troncoso, C., Diaz, C.: Engineering privacy by design. Computers, Privacy & Data Protection **14**(3), 25 (2011)
10. McManus, J.: Security by design: teaching secure software design and development techniques. J. Comput. Sci. Coll. **33**(3), 75–82 (2018)
11. Felzmann, H., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A.: Towards transparency by design for artificial intelligence. Sci

Eng Ethics **26**(6), 3333–3361 (2020). https://doi.org/10.1007/s11948-020-00276-4
12. Kelty, C.M.: Beyond implications and applications: the story of 'Safety by design.' NanoEthics **3**(2), 79–96 (2009). https://doi.org/10.1007/s11569-009-0066-y
13. Deogun, D., Johnsson, D. B., Sawano, D.: Secure by Design. Manning Publications (2019)
14. Rocha, C.S., Antunes, P., Partidário, P.: Design for sustainability models: a multiperspective review. J. Clean. Prod. **234**, 1428–1445 (2019)
15. Allison, I.: "Organizational factors shaping software process improvement in small-medium sized software teams: A multi-case analysis", in. Seventh International Conference on the Quality of Information and Communications Technology **2010**, 418–423 (2010)
16. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
17. High-Level Expert Group on Artificial Intelligence, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe's digital future," European Commission, Jul. 2020. Accessed: Mar. 07, 2023. [Online]. Available: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342
18. Gunning, D.: "Explainable artificial intelligence (xai)," Defense advanced research projects agency (DARPA), nd Web, **2**(2): 1 (2017)
19. Nasim, S. F., Ali, M. R., Kulsoom, U.: "Artificial intelligence incidents & ethics a narrative review." Int. J. Technol. Innov. Manag (IJTIM) (2022). https://doi.org/10.54489/ijtim.v2i2.80
20. Liu, D., Oberman, H. I., Muñoz, J., Hoogland, J., Debray. T.: 'Quality Control, Data Cleaning, Imputation'. ArXiv Preprint ArXiv:2110.15877, (2021)
21. Raji, I. D. et al.: "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in Proceedings of the 2020 conference on fairness, accountability, and transparency, pp. 33–44 (2020)
22. Mitchell, M. et al.: "Model cards for model reporting," in Proceedings of the conference on fairness, accountability, and transparency, pp. 220–229 (2019)
23. Gebru, T., et al.: Datasheets for datasets. Commun. ACM **64**(12), 86–92 (2021)
24. Sokol, K., Flach, P.: "Explainability fact sheets: a framework for systematic assessment of explainable approaches," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 56–67 (2020)
25. Rabelo, J. de H. et al.: "Knowledge Management and Organizational Culture in a Software Organization – A Case Study," in 2015 IEEE/ACM 8th International Workshop on Cooperative and Human Aspects of Software Engineering, 89–92. (2015) https://doi.org/10.1109/CHASE.2015.27.
26. Bannon, L. J.: "From Human Factors to Human Actors: The Role of Psychology and Human-Computer Interaction Studies in System Design," in Readings in Human–Computer Interaction, Baecker RM, Grudin J, Buxton WAS, Greenberg S, (Eds). Morgan Kaufmann, pp. 205–214. (1995) https://doi.org/10.1016/B978-0-08-051574-8.50024-8.
27. Wei, M., Zhou, Z.: 'AI Ethics Issues in Real World: Evidence from AI Incident Database'. ArXiv Preprint ArXiv:2206.07635, (2022)
28. Štefanišinová, N., Jakuš Muthová, N., Štrangfeldová, J., Šulajová, K.: Implementation and application of artificial intelligence in selected public services. Hrvatska i komparativna javna uprava: časopis za teoriju i praksu javne uprave **21**(4), 601–622 (2021)
29. "XAI - An eXplainability toolbox for machine learning." The Institute for Ethical Machine Learning, Feb. 21, 2023. Accessed: Feb. 22, 2023. [Online]. Available: https://github.com/EthicalML/xai
30. Mantelero, A.: Beyond data: human rights, ethical and social impact assessment in AI. Springer Nature (2022). https://doi.org/10.1007/978-94-6265-531-7
31. Nitta, I., Ohashi, K., Shiga, S., Onodera, S.: "AI Ethics Impact Assessment based on Requirement Engineering," in 2022 IEEE 30th International Requirements Engineering Conference Workshops (REW), pp. 152–161. (2022) https://doi.org/10.1109/REW56159.2022.00037
32. Brey, P., Dainow, B.: Ethics by design and ethics of use in AI and robotics. SIENNA (2021). https://sienna-project.eu/digitalAssets/915/c_915554-l_1-k_sienna-ethics-by-design-and-ethics-of-use.pdf
33. Brey, P.: Research ethics guidelines for artificial intelligence. In: Resseguier, A., Brey, P., Dainow, B., Drozdzewska, A. (eds.) Annex 4 of SIENNA D5.4: Multi-stakeholder Strategy and Practical Tools for Ethical AI and Robotics, pp. 100–126 (2021). https://doi.org/10.5281/zenodo.5536176
34. European Commission, "How to complete your ethics self-assessment (2.0)," Jul. 2021. [Online]. Available: https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf
35. McLennan, S et al.: "An embedded ethics approach for AI development," Nat Mach Intell **2**(9), Art. no. 9, (2020), https://doi.org/10.1038/s42256-020-0214-1.
36. Kearns, M., Roth, A.: The ethical algorithm: the science of socially aware algorithm design. Oxford University Press (2019)
37. Pessach, D., Shmueli, E.: "Algorithmic fairness," arXiv preprint arXiv:2001.09784, (2020)
38. Friedman, B., Hendry, D. G.: Value sensitive design: Shaping technology with moral imagination. Mit Press, (2019)
39. Systems and Software Engineering Standards Committee, "IEEE Standard Model Process for Addressing Ethical Concerns During System Design: IEEE Standard 7000–2021," (2021)