From Principles to Practice

An interdisciplinary framework to operationalise AI ethics


VDE


BertelsmannStiftung


H L R S
High Performance Computing Center | Stuttgart


KIT
Karlsruhe Institute of Technology


TECHNISCHE
UNIVERSITÄT
DARMSTADT


TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN


EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN
INTERNATIONAL CENTER
FOR ETHICS IN THE SCIENCES
AND HUMANITIES (IZEW)


iRights.Lab

From Principles to Practice

An interdisciplinary framework to operationalise AI ethics

CONTENTS

EXECUTIVE SUMMARY

Artificial intelligence (AI) increasingly pervades all areas of life. To seize the opportunities this technology offers society, while limiting its risks and ensuring citizen protection, different stakeholders have presented guidelines for AI ethics. Nearly all of them consider similar values to be crucial and a minimum requirement for "ethically sound" AI applications – including privacy, reliability and transparency. However, how organisations that develop and deploy AI systems should implement these precepts remains unclear. This lack of specific and verifiable principles endangers the effectiveness and enforceability of ethics guidelines. To bridge this gap, this paper proposes a framework specifically designed to bring ethical principles into actionable practice when designing, implementing and evaluating AI systems.

We have prepared this report as experts in spheres ranging from computer science, philosophy, and technology impact assessment via physics and engineering to social sciences, and we work together as the AI Ethics Impact Group (AIEI Group). Our paper offers concrete guidance to decision-makers in organisations developing and using AI on how to incorporate values into algorithmic decision-making, and how to measure the fulfilment of values using criteria, observables and indicators combined with a context-dependent risk assessment. It thus presents practical ways of monitoring ethically relevant system characteristics as a basis for policymakers, regulators, oversight bodies, watchdog organisations and standards development organisations. So this framework is for working towards better control, oversight and comparability of different AI systems, and also forms a basis for informed choices by citizens and consumers.

The report does so in four steps:
In chapter one, we present the three main challenges for the practical implementation of AI ethics: (1) **the context-dependency of realising ethical values**, (2) **the socio-technical nature of AI usage** and (3) **the different requirements of different stakeholders concerning the 'ease of use' of ethics frameworks**. We also explain how our approach addresses these three challenges and show how different stakeholders can make use of the framework.

In chapter two, we present the VCIO model (values, criteria, indicators, and observables) for the operationalisation and measurement of otherwise abstract principles and demonstrate the functioning of the model for the values of transparency, justice and accountability. Here, we also propose context-independent labelling of AI systems, based on the VCIO model and inspired by the energy efficiency label. This labelling approach is unique in the field of AI ethics at the time of writing.

For the proposed AI Ethics Label, we carefully suggest six values, namely justice, environmental sustainability, accountability, transparency, privacy, and reliability, based on contemporary discourse and operability.

Chapter three introduces the risk matrix, a two-dimensional approach for handling the ethical challenges of AI, which enables the classification of application contexts. Our method simplifies the classification process without abstracting too much from the given complexity of an AI system's operational context. Decisive factors in assessing whether an AI system could have societal effects are the intensity of the system's potential harm and the dependence of the affected person(s) on the respective decision. This analysis results in five classes which correspond to increasing regulatory requirements, i.e. from class 0 that does not require considerations in AI ethics to class 4 in cases where no algorithmic decision-making system should be applied.

Chapter four then reiterates how these different approaches come together. We also make concrete propositions to different stakeholders concerning the practical use of the framework, while highlighting open questions that require a response if we ultimately want to put ethical principles into practice. The report does not have all the answers but provides valuable concepts for advancing the discussion among system developers, users and regulators.

Coming together as AI Ethics Impact Group, led by VDE Association for Electrical, Electronic & Information Technologies and Bertelsmann Stiftung and presenting our findings here, we hope to contribute to work on answering these open questions, to refine conceptual ideas to support harmonisation efforts, and to initiate interdisciplinary networks and activities.

We look forward to continuing the conversation.

AIEI Group
Artificial Intelligence Ethics Impact Group

Dr Sebastian Hallensleben
Carla Hustedt

Lajla Fetic
Torsten Fleischer
Paul Grünke
Dr Thilo Hagendorff
Marc Hauer, Andreas Hauschke
PD Dr Jessica Heesen
Michael Herrmann
Prof. Dr Rafaela Hillerbrand
Prof. Emeritus Christoph Hubig
Dr Andreas Kaminski
Tobias Krafft,
Dr Wulf Loh
Philipp Otto
Michael Puntschuh

1 INTRODUCTION

With the increasing use of artificial intelligence in all areas of life, often in the form of algorithmic decision-making systems (ADM), discussions about AI ethics are omnipresent, be it in the scientific community, in organisations developing or using AI, in standard-setting bodies and regulatory institutions, or civil society. Consequently, national, as well as European institutions, are in the process of formulating frameworks for algorithmic decision-making in which ethics and the underlying principles and values are critical aspects.

To date, well over a hundred different AI ethics guidelines have been published.[1] Nearly all of them mention values such as privacy, fairness or non-discrimination, transparency, safety, and accountability. These seem to be considered the minimum requirements for building and using AI applications that could be deemed ethical. The categories also form the basis for the so-called European approach to artificial intelligence, as can be seen in the various initiatives taken on the EU level.[2] However, the way in which organisations that develop and deploy such applications should implement these precepts is unclear.

Implementation challenge

The lack of specific and verifiable principles endangers the effectiveness of ethical guidelines. It creates uncertainty and raises concerns about new red tape among organisations developing AI. Lack of specificity impedes the work of oversight bodies and watchdog organisations that cannot measure their implementation if principles remain vague and thereby hinders the enforceability of the guidelines. It leaves policymakers wondering how to formulate regulation that protects European values and citizens without inadvertently hindering growth and innovation.

The solution:

Solution framework

In bringing together this interdisciplinary group, we aimed to fill this gap by introducing a framework that demonstrates how to put ethical principles into AI practice. This report thus advances the discussion through multiple contributions:

- We present the so-called VCIO (Values, Criteria, Indicators, Observables) model, an approach for the specification and operationalisation of values. This is necessary to make ethical principles practicable, comparable and measurable. We also demonstrate different ways of dealing with conflicts between values.

1 These include the April 2019 report of the European High-Level Expert Group on AI, the AI4People framework (Floridi 2018), the Beijing AI Principles (2019), or the OECD Recommendation of the Council on Artificial Intelligence (Organisation for Economic Co-operation and Development (2019) to name a few. For an overview, see Fjeld et al. (2020), Jobin et al. (2019), or Hagendorff (2020).

2 See the Guidelines of the European High-Level Expert Group.

- We offer practical examples for applying the VCIO model to selected values such as transparency, accountability and justice. By operationalising selected principles in this paper, we also identify key challenges and open questions relevant for standard-setting and regulation, which should form the basis for future work in AI ethics.

- We propose ratings for AI ethics, as illustrated in the AI Ethics Label, inspired by the energy efficiency label. It can offer orientation to developers trying to create ethically sound AI systems, increase transparency and comparability of products for users, and provide a basis for better oversight by policymakers, regulators, standards developing associations, and watchdog organisations.

- We present the risk matrix, a two-dimensional model for the classification of different application contexts of AI systems.

- We show how these different approaches, i.e. the VCIO model, ethics rating and risk matrix, work together and how the framework can be used by organisations developing or using the systems as well as by policymakers, regulators, oversight bodies, watchdog, and standard-setting organisations.

Viable at multiple levels

Our work and this paper build on previous and ongoing discussions, and we conducted it in a science-based, interdisciplinary and participatory manner. The method is viable on the national, European and international level. It aims to form a basis for the work of policymakers, regulators, oversight bodies, watchdog- and standard-setting organisations while offering orientation for self-commitment by individuals and entities developing and deploying AI systems at a time where no official standard or regulation yet exists.

Defining AI

What do we mean by artificial intelligence? AI is not straightforward to define, so instead of a definition, we use the following orientation for the scope considered here.

In its current and most common usage, the term artificial intelligence refers to non-symbolic machine learning techniques or the science of getting computers to act without being explicitly programmed. One of the best-known techniques is deep learning. In general, we categorise machine learning techniques as supervised, unsupervised or reinforced learning. Supervised learning forms the basis of many algorithmic decision-making systems (ADM). To take autonomous decisions, those ADM systems derive information from considerable datasets (big data). Other typical areas of application in machine learning include machine vision, natural language processing, and robotics, to name a few.

In this framework, we further extend our definition by all technical approaches that rely on symbolic AI, i.e. systems based on rules and that pursue weak AI and that comprise techniques of machine learning, whereby this includes not only deep learning, i.e. neural networks but support-vector machines, typically used for regression methods, and also Bayes algorithms.

To conclude, rather than looking on specific technical features of AI, we look at all technical systems that whether non-symbolic or symbolic might have an impact on humanity through (partial) automation of decision-making processes.

1.1 Challenges of practically implementing AI ethics

Values that form the basis for AI ethics are open to different interpretations and may vary depending on an AI system's application context. Additionally, the complexity and a high number of stakeholders involved in the development and implementation of AI systems raises further challenges for the effective enforcement of ethical principles. Consequently, for any AI ethics framework to have an impact in practice, the implementation needs to address three main challenges:

Context-dependence

(1) The realisation of values depends on the field of application and cultural context: Values for AI systems must be fleshed out through contextualised interpretations and their application to situations. So how we implement and prioritise values such as justice (here includes fairness or non-discrimination) and transparency in practice, depends to some extent on the field of application and the cultural context an AI system operates in. A system used in the justice sector must necessarily exhibit higher levels of privacy and fairness than a system used in the organisation of industrial production. For application in the medical sector, reliability could be considered the most critical value. Besides, there are often conflicts between such values, typically reframed as trade-offs, so for example, the more transparent a system is, the lesser privacy it may protect. Also in different cultural contexts, values are given different priorities. One culture may prefer one value over another.

▸▸ *Any framework for the practical implementation of AI ethics must take these differences for the realisation of values (considering value fulfilment), potential value conflicts, and heterogeneity of application contexts into account.*

Socio-technical nature of AI

(2) Multiple factors in an AI system's development and implementation influence its impact: AI systems are socio-technical systems. Their societal impact depends not only on the technology (data and algorithms) but also on the system's underlying goals and on the way an AI is embedded in an organisational structure. The implementation of values such as justice and transparency requires multiple measures throughout the complex development and implementation process of AI systems. For example, transparency might depend on the technical explainability of an AI system, which lies in the hands of system developers, but also requires active communication and explanation of algorithmic decision-making processes in organisations using the AI system.

▸▸ *Actionable frameworks for AI ethics need to consider the complex development and implementation process, the socio-technical nature of AI systems, and the responsibilities of system developers and users derived thereof.*

Ease of use

(3) 'Ease of use' of an AI ethics framework means different things for different stakeholders: Due to the socio-technical nature of AI systems, frameworks for the practical implementation of AI ethics need to provide tools that take into account the different roles played by system developers and system users in providing necessary measures. Such frameworks also need to ease external scrutiny over the implementation of these measures (enforceability).

However, the different stakeholders involved in the development, implementation and evaluation process have different requirements when it comes to the usability of frameworks for ethical considerations. Vendors of AI systems need an approach that makes the implementation of such principles as easy as possible. AI developers who are generally equipped to deal with technical challenges but not ethical dilemmas need straightforward guidance. Organisations using AI systems need tools that provide simple comparability of different offers when procuring the technical systems as well as guidance for embedding these in their organisational structure. And Europeans, both as citizens and consumers, expect AI ethics to be communicated in a way that is immediately comprehensible and allows them to assess the quality of the systems that affect their lives.

▸▸ ***As all these stakeholders have different levels of responsibility for technical and ethical questions, it is crucial for any AI ethics framework to simplify but not oversimplify and to provide guidance appropriate to each stakeholder's requirement.***

Combining system and process perspectives

For our approach, we combine a system perspective that defines specific ethical requirements on AI systems themselves with a process perspective that specifies requirements for the design and implementation processes. Both perspectives have their strengths and weaknesses:

In the system perspective, a value such as transparency is easily verifiable and, therefore, allows for good comparability of AI systems and enforcement of values. At the same time, a process perspective is better suited for handling the socio-technical nature of AI and situations where value realisations and prioritisation are controversial and context-dependent.
A pure process perspective, however, may invite "fig leaf behaviour" by AI vendors, especially large companies with well-staffed compliance departments. It also tends to place ethically sensitive decisions with societal relevance into internal and often non-transparent committees and relies on the good intentions of stakeholders which, especially from a regulatory perspective, cannot be taken for granted. So overall, for a practical approach to handling AI ethics, it is necessary to use a "best of both worlds" method that draws on both system and process perspectives.

1.2 Multimethod framework as solution

Framework addresses all main challenges

There is no lack of proposals for handling ethical aspects of AI. Many of these, however, do not adequately address the challenges of context-dependency, the socio-technical nature of AI systems and the different requirements for the 'ease of use' of system developers, users, oversights bodies, policymakers, and consumers. We, therefore, introduce a framework that focuses on the three main challenges we have identified:

(1) Our answer to the challenge of context-dependency: Combination of a context-independent ethics rating and a classification approach

In a first step, our framework introduces an approach for the rating of ethically relevant characteristics of an AI system (e.g. with regards to justice, accountability, transparency) independent of the system's application context (see VCIO approach, chapter 2). In a separate step, it introduces a classification of different AI application contexts, based on the risk they pose for the individuals affected and society overall (see risk matrix, chapter 3). Our framework describes AI systems without also deciding what is acceptable and what is not, and thus leaves those judgements in the hands of regulators and users.

(2) Our answer to the challenge of the socio-technical nature of AI systems: Specification of general principles both in terms of technical system requirements as well as implementation requirements

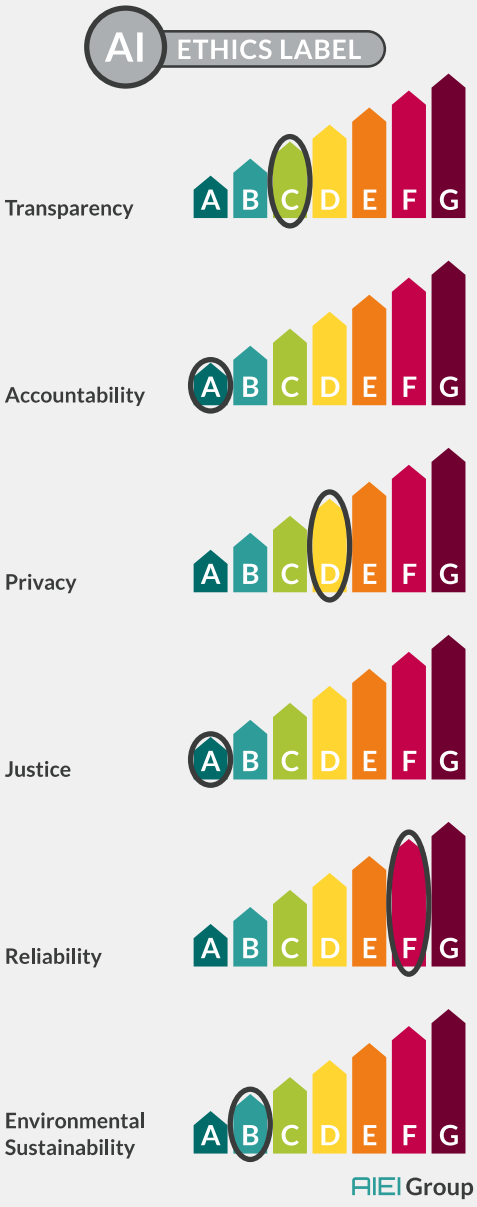
The measurements and observables for specific values presented in chapter 2 include both requirements for the technical system (targeting system developers) as well as requirements for the implementation process of the system (targeting system users). Depending on who uses our framework at which point in development and implementation, the focus can be on either the system or process requirements. For policymakers, regulators, oversight bodies and watchdog organisations as well as consumers and citizens affected, both types of requirements are relevant to assess to what extent a system is suitable for a particular application area or not.

(3) Our answer to the challenge of different needs of stakeholders concerning 'ease of use': Introduction of a nuanced labelling approach

In chapter 2.3, our framework introduces the idea of a nuanced AI Ethics Label, inspired by the well-known energy efficiency label. Labels indicating the characteristics of a product have become well-established in many industries and have proven useful and acceptable to consumers, industry and regulators alike. It is applicable as self-commitment but can be used for stricter regulation as well, and therefore is workable for all stakeholders involved. For example, it can become both a template for the work of regulatory bodies commissioned with the enforcement of regulation and provide orientation to AI developers and users and citizens and consumers.

Taking the energy efficiency label as a guide, a label showing a rating of an AI system's ethical characteristics could then look as follows:

FIGURE 1 **The AI Ethics Label with six selected values**



1.3 Handling AI ethics in practice

Orientation for stakeholders to bring AI ethics into practice

Taken together, our approach for the operationalisation of general principles (VCIO), the context-independent rating of ethical characteristics, the proposal of the introduction of an AI ethics label and the classification of different application contexts through a risk matrix provides a framework for bringing AI ethics from principles to practice. While we explain and discuss the framework's elements in more depth in chapters 2 and 3, we here give an overview of how different stakeholders can use the approaches:

- An organisation planning to use an AI system for a specific application follows a simple initial checklist (e.g. drawing on the experience from HLEG pilots) to determine whether the application is ethically-sensitive. If an application is rated as non-ethically sensitive, the process ends at this stage. There is, for example, no need to consider ethical requirements in purely industrial applications, that do not affect people's lives in a significant manner.

  However, if the triage indicates that there are ethical issues to consider, then the organisation performs a full assessment of the application context using the risk matrix. If no regulation or official standard for their application field exists, they can then use the VCIO approach to concretize general ethical principles for the use of AI as a basis for self-commitments.

- Procurement departments (both in the private and public sector) use ethics rating and risk matrix to create clear specifications for the AI systems they plan to use. They also benefit from market transparency through the AI ethics label. In many cases, automated filtering for the desired ethics rating is possible, e.g. when reviewing product catalogues or visiting AI-backed websites.

- Similarly, manufacturers of AI systems can consider the range of expected applications using the risk matrix and decide whether to market an AI system only for applications without ethical sensitivity, or also for higher risk classes. In the latter case, they may gain market advantage by achieving a high ethics rating recognised worldwide for their products. This applies in both B2C and B2B settings.

- Regulators can use the combination of the risk matrix and ethics rating to specify requirements for different application contexts and to avoid over-regulation of application fields that do not pose any significant ethical challenges. For application fields that are classified in one of the higher risk levels, they may demand that an AI system (1) must carry an ethics label that shows the rating for values such as transparency, robustness, or justice and (2) satisfy minimum levels within the rating.

- Consumers use the ethics rating to compare AI products and services and make informed decisions about what is acceptable to them and/or worth investing in. Consumers are alerted to ethically sensitive applications through the risk matrix. Moreover, consumers can trust that minimum regulatory requirements protect them.

2 VALUES, CRITERIA, INDICATORS, OBSERVABLES (VCIO) AND THE AI ETHICS LABEL IN DETAIL

Why VCIO?

The VCIO model distinguishes and combines the four concepts of values, criteria, indicators and observables for the evaluation of AI.[3] The question is, however, why do we need criteria, indicators, and observables, in short, the CIO-part of the VCIO approach? As values are abstract, often in conflict with each other, and do not include means to evaluate their implementation, it is essential to have other components to fulfil these tasks. This is where the criteria, indicators and observables of the VCIO approach come into play.

For example, the demand that algorithms should not discriminate finds consensus; the debate, however, begins with the question of what is understood by discrimination (justice), how to check whether it exists, and how to deal with conflicts between different values.

Similarly, consider sustainability as a value. The fact that technologies should be sustainable is unlikely to cause any contradiction. The dispute usually starts with what constitutes sustainability.

The VCIO approach, therefore, fulfils three tasks:

1) clarifies what is meant by a particular value (value definition)
2) explains in a comprehensible manner how to check or observe whether or to what extent a technical system fulfils or violates a value (measurement)
3) acknowledges the existence of value conflicts and explains how to deal with these conflicts depending on the application context (balancing).

To practically implement AI ethics, the VCIO approach operates on its four levels as follows:

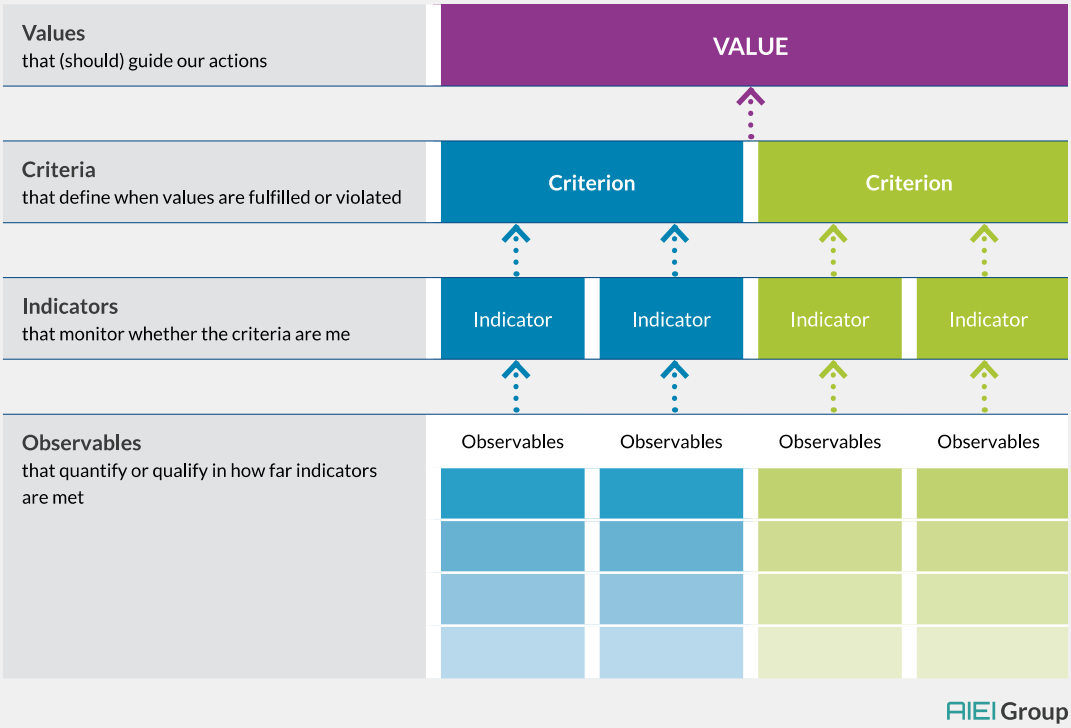
VCIO: values, criteria, indicators, and observables

Values formulate a general ethical concern, something that should guide our actions. They are defined at the highest level (as justice or transparency, for example). To verify whether an algorithm fulfils or violates specific values, we must specify Criteria that define the fulfilment or violation of the respective value. Since it is usually not possible to directly observe whether a criterion is met, we need Indicators (as a specific type of sign) to monitor this. Indicators relate criteria on the one hand with Observables on the other.[4]

3 For the foundation of this approach, see Hubig 2016.

4 Indicators cannot be logically deduced from criteria, but rather form necessary conditions for the fulfilment of criteria on an abstract level. They are posited and set through deliberative processes. As necessary conditions, they cannot be weighed against or substituted for each other. By focusing on the respective observables (in the form of quantitative or qualitative values), we gain insight into the implementation status of the indicator in question. These values can be related across indicators with regard to amplification, compensation or "killer" effects. The latter refers to an exceeding/undercutting of specific negotiable boundary values.

FIGURE 2 **The VCIO model**



Each level's normative load

The four hierarchical levels provided by values, criteria, indicators and observables are closely linked, where the fulfilment of the higher level depends on the lower level. However, it is not possible to derive the lower levels from the higher ones in a straightforward, i.e. deductive way. Instead, the normative load runs through all four levels and requires new deliberations at all levels, in the course of which the particular instances must be negotiated in detail.[5]

Note that typically, several indicators are required to evaluate the fulfilment of a criterion; however, we can also use an indicator to assess the fulfilment of various criteria. As there are no deductive relationships between values, criteria, indicators, and observables, as a rule, at each stage of their determination, normative decisions should be made in a scientific and technically informed context.

To go back to the example of sustainability: when looking at 'sustainability' as a value, we may use 'resource depletion' as a criterion. Naturally, the choice of value is already normative to an extent, i.e. the implicit assumption that sustainability is relevant in an ethical sense. So the choice of indicators, as well as zooming in on some observables, depends on available data and has normative aspects. Therefore, the decision for different

5 This is also known as "indicator politics" (see Hubig 2016, cf. van der Poel 2013). Also, it is not possible to logically deduce criteria or indicators from values but to argue for these in deliberative processes.

criteria, indicators or observables is as normative as it is for specific values. It is not possible to logically deduce criteria or indicators[6] from values but to argue for these in deliberative processes.

For further illustration, consider the problem of applying the value 'sustainability' to a lake as outlined by Hubig (2016, p. 6–7). Taking the derived value 'water quality' and we find two conflicting criteria: 'drinking water quality' and 'quality for the food chain of fish populations'. The populace using the lake as a drinking water reservoir stick to the first criterion, but fisher will most likely prefer the latter. The corresponding indicator 'phosphate level' demonstrates the conflict between both criteria. A lower phosphate level is better for drinking water quality than for the fish population.

2.1 How to apply VCIO to AI ethics: Three illustrated examples

Operationalising values with indicators, observables and potential value conflicts

In the following, we illustrate how to apply the VCIO model by focusing on three values, namely transparency, justice, and accountability.[7] The findings from the Algo.Rules project, an initiative by the Bertelsmann Stiftung and the iRights.Lab, have been essential for the development of the framework and this chapter in particular.

Each of the illustrations shows one value with corresponding criteria, indicators, and observables suggested. In defining these, we combine the system with process requirements. The observables show the different levels on which we can observe the indicator in a given system. Colour coding sorts the gradations within observables from dark (best, high) to light (worst, low). A higher level within the observables corresponds with a higher rating of the overall value.

Next to the values, criteria, indicators and observables, we added comments for the value overall as well as for specific criteria, indicators or observables to explain why we have chosen individual specifications and how they could be used or further developed.

We are aware that the examples given in the following and the illustrations here are neither comprehensive nor detailed enough for direct application in the field. Instead, the following illustrations are a suggestion and work-in-progress, as they would need to be further negotiated, specified and reflected on as part of any labelling or harmonising process. However, they may serve as guidelines for the kind of processes and discussions regulators or users may undergo in practice.

6 Indicators thus form necessary conditions for the fulfilment of criteria on an abstract level.

7 More information on the Algo.Rules, 9 principles for the ethical development of algorithmic systems, can be found here: https://algorules.org/en/home

Dealing with conflicting values

The VCIO-approach allows measuring the fulfilment of values using criteria, observables and indicators. However, conflicts between values and associated indicators may arise. A conflict of values emerges when a value and its indicators can't be valid only to the extent of not violating another value. Here, the VCIO model's basic principle of making values measurable cannot be achieved. Instead, value conflicts are a component of the approach with the assessment and also resolving such conflicts depending on the application context and the regulator's or user's perspective on it.

If conflicts exist, values and indicators that meet an application's requirements must be hierarchised (see Hubig 2007, Section 3.2). The resolve of conflicts of values can thus happen in two places through different stakeholders:

(1) The regulator has to hierarchise the values concerning the context of an AI application (also see risk matrix), e.g. in negotiation with all parties affected.
(2) Users can solve the value conflicts for themselves according to their preferences, either not taking into account the classification of application context or going beyond the minimum requirements.

To hierarchise or manage values conflicts, there are the following two strategies from the tradition of wisdom ethics:

1) Bottom-up: The starting point here is a problem's urgency. It arises from the danger that if the issue remains unsolved, subsequent action at a whole would be difficult. (Aristotle's considerations on equity aim at determining when the violation of values is permitted and appropriate in a concrete case.)
2) Top-down: Certain values are to be regarded as higher-level because they affect further action beyond the situation. To decide this, look at option values. These concern subsequent possibilities or scope for action. Depending on which path of action we choose, this reduces or increases our future options for action. Path dependencies or "constraints" show that option values can be violated.

There are also legacy values, which concern our ability to relate to our decisions and actions in an evaluative manner. That legacy values can be violated becomes evident in the face of informal techniques where we find it difficult to evaluate our decisions and those of the technique.

In practice, where values such as privacy, reliability or justice come into conflict with each other, option and legacy values can act as arbitrator values. We then assess the extent in which a resolution of the conflict in favour of one value (such as reliability) at the expense of another (such as privacy) reduces or increases our scope for action or affects our ability to place ourselves in an evaluative relationship to our action's consequences.

Cascading Values

Transparency

Transparency as explainability and interpretability

For this rating, the value of transparency is understood as explainability and interpretability of the algorithmic system, including the model and data used. The question is here how or in how far transparency is being achieved. Transparency, therefore, refers to disclosing the data's origin and properties of the AI model in use as well as access to and comprehensibility of the information disclosed. In this sense, we aim for transparency in both the general operating principle and each output of the AI system. Transparency furthermore must be tailored to the requirements of the target groups such as users and persons affected, i.e. the system must be comprehensible to them. **2.1.1 Applying the VCIO approach to transparency as a value (page 20/21)**

Justice

Justice with aspects of algorithmic fairness and inclusion

The criteria subsumed under the value of justice in this example pertain to classic aspects of algorithmic fairness such as bias prevention and assessment but emphasise a process perspective to include a broader set of ethical considerations. These aspects are, for example, inclusion, represented by criteria such as participatory procedures, or social justice considerations, and a criterion for the assessment of trade-offs generated by the employment of the AI system in question. In this sense, justice refers to a broader set of ethical considerations than the often-used term fairness, which mostly focuses on algorithmic outcomes themselves. **2.1.2 Applying the VCIO approach to justice as a value (page 22/23)**

Accountability

Accountability refers to questions of assigning responsibility

The value of accountability refers to problems that arise in connection with the complex allocation or clarification of responsibility relationships in the use of AI. The various dimensions of accountability range from retrospective to prospective organisational measures for assigning responsibilities. They also include technical means or specific ways of dealing with organisational and technical errors. **2.1.3 Applying the VCIO approach to accountability as a value (page 24/25)**

2.1.1 Applying the VCIO approach to transparency as a value

Value	TRANSPARENCY					
Criteria	Disclosure of origin of data sets			Disclosure of properties of algorithm/model used		
Indicators	Is the data's origin documented?	Is it plausible for each purpose, which data is being used?	Are the training data set's characteristics documented and disclosed? Are the corresponding data sheets comprehensive?	Has the model in question been tested and used before?	Is it possible to inspect the model so far that potential weaknesses can be discovered?	Taking into account efficiency and accuracy, has the simplest and most intelligible model been used?[1]
Observables	Yes, comprehensive logging of all training and operating data, version control of data sets etc.[2]	Yes, the use of data and the individual application are intelligible	Yes and the data sheets are comprehensive	Yes, the model is widely used and tested both in theory and practice[3]	Yes, the model can easily be inspected and tested	Yes, the model has been evaluated and the most intelligible model has been used
	Yes, logging and version control through an intermediary (e.g. data supplier)	Yes, it is intelligible on an abstract, not case specific level, which data is being used	Yes, but (some) data sheets contain few or missing information	Yes, the model is known and tested in either theory or practice	Yes, but the model can only be tested by certain people due to non-disclosure	No, but the model was evaluated regarding interpretability and this evaluation is disclosed to the public
	No logging; data used is not controlled or documented in any way	No, but a summary on data usage is available	No	Yes, the model is known to some experts but has not been tested yet	No	No, the model has not been evaluated
		No		No, the model has been developed recently		

[1] This indicator would require further specification regarding the balance between using an efficient and accurate model and using a model which is technically simple and thus naturally easier to comprehend and follow.
[2] This observable could include further levels of logging and documentation of data sets.
[3] This observable could help to determine the levels needed in other observables: If the model has been widely used and tested, it might not require additional testing.

TRANSPARENCY						Value
Accessibility						Criteria
Are the modes of interpretability target-group-specific and have been developed with the target groups?	Who has access to information about data sets and the algorithm/model used?	Is the operating principle comprehensible and interpretable?	Are the modes of interpretability in their target-group-specific form intelligible for the target groups?	Are the hyperparameters (parameters of learning methods) accessible?	Has a mediating authority been established to settle and regulate transparency conflicts?	Indicators
Yes	Everyone	Yes, the model itself is directly comprehensible	Yes, the modes of interpretability have been tested with target groups for intelligibility	Yes, to everyone	Yes, a competent authority has been established	Observables
Yes, but without participation of the target groups	All people directly affected	Yes, the modes of interpretability are provided with the model itself				
Yes, but the modes or interpretability are only specific for one target group	Only information and trust intermediaries (regulators, watchdogs, research, courts)	No, the modes of interpretability can only be used post hoc by experts	Yes, target groups can complain or ask if they do not understand a mode of interpretability	Yes, but only to information and trust intermediaries (regulators, watchdogs, researchers, courts)	Yes, a competent authority has been established but its powers are limited	
No, the modes of interpretability[4] are not target-group-specific	Nobody	No, the modes of interpretability need to be adjusted to the individual model and use by experts				
		No, but the model is theoretically comprehensible	No	No	No	
		No, there are no known modes of interpretability				

[4] "Modes of interpretability" refers to different methods to ensure or increase interpretability (use of simple model, explanations of data and model used, etc.).

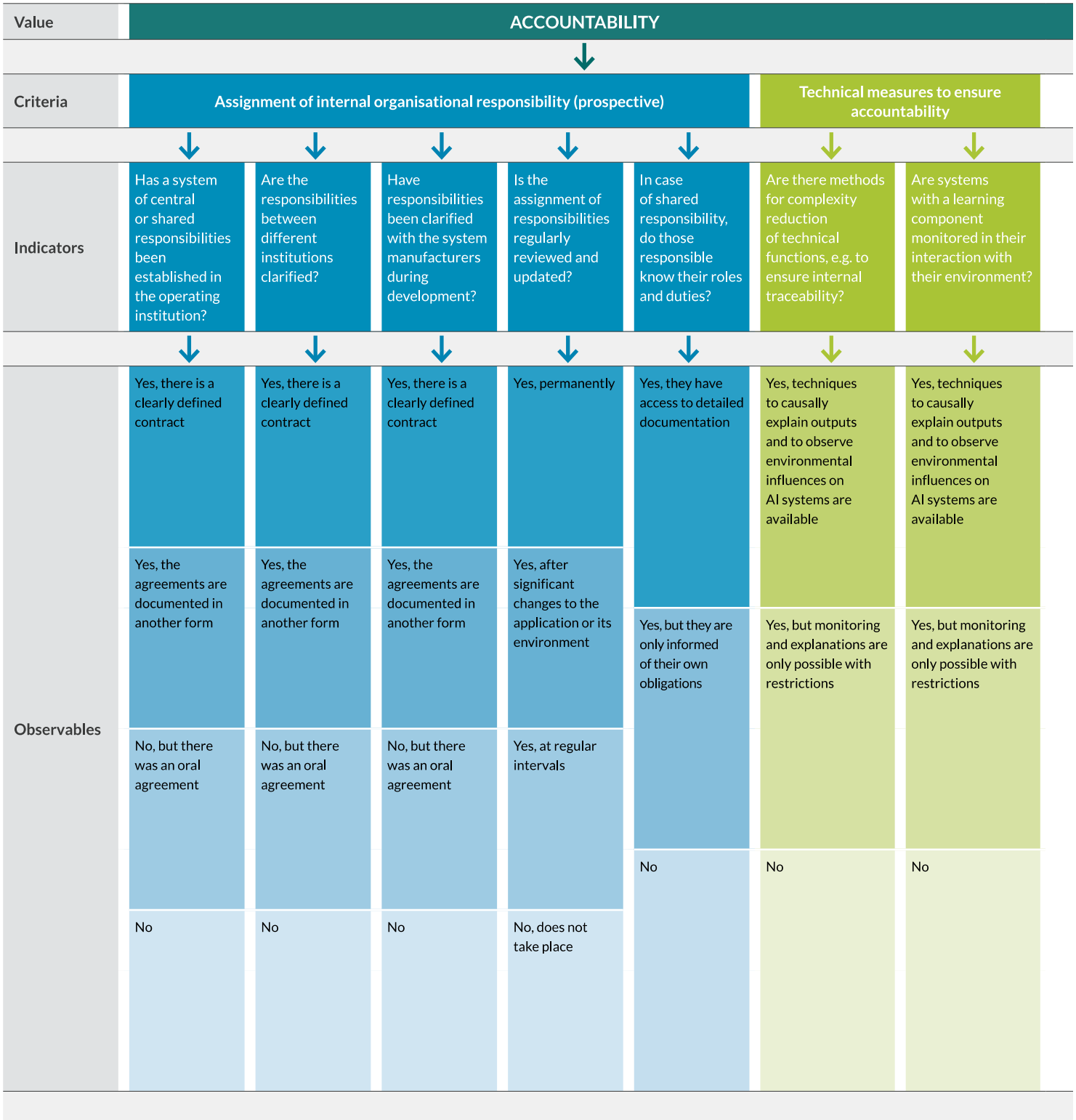
2.1.2 Applying the VCIO approach to justice as a value

Value	JUSTICE								
Criteria	Identifying and assessing trade-offs	Assessment of different sources of potential biases to ensure fairness[1]							Social justice considerations
Indicators	Have trade-offs been identified and assessed?	Has the training data been analysed for potential biases?	Has the input design (sensors, user interface) and input data been reviewed for potential biases?	Have the requirements, goals and task definitions been examined for implicit and explicit discriminatory effects?	Were possible self-reinforcing processes considered?	Has due care been taken with regard to discriminatory effects caused by the design of the data output?	Have the applied methods (e.g. categorisation) been evaluated for potential biases and discriminatory effects?	Is a special checking procedure for possible proxies of sensitive data in place? Is the collection of proxies avoided?	Have the working conditions, e.g. data labelling procedures, been evaluated?[3]
Observables	Yes, with the help of a regular external technology impact assessment	Yes, demographic parity, equality of odds and opportunities are ensured	Yes, review on a regular basis	Yes, and continual reviews are conducted	Yes, periodically	Yes, and the output data design is periodically reviewed	Yes	Yes, continual checks	Yes, employing external evaluation mechanisms
	Yes, with the help of an external technology impact assessment, but only once	Only limited assessment	Yes, but only once	Yes, after changes of the application or its environment	Yes, but only once	Yes, but only during the development/ implementation process		Yes, checks are made after changes to the application	Yes, but only internal assessment
	Yes, but only internal impact assessment			Yes, periodically			No	Yes, periodically	
		No	No		No[2]	No			No
	No			No				No	

[1] This includes biases produced by the algorithm as well as existing societal biases that are proliferated and perpetuated by the algorithmic system.
[2] The periodization is only applicable to a system that continues to adapt through day-to-day input data.
[3] While the ethics rating does not necessarily reflect the working conditions of the system developers and operators, this criterion takes into account various forms of so-called click work. The latter is essential for data annotation required for training and assessment of ADM systems.

Value	Criteria	Indicators	Observables
JUSTICE	Detection and prevention of biases to ensure fairness	Is there an external investigation of error sources?	Yes, by an independent institution
			Yes, by a trade association or another related institution
			Yes, through an internal department
			Yes, by the AI developing team itself
			No
		Are mechanisms in place to provide access to data/ processes for third-party evaluation?	Yes, public access
			Yes, but access will only be granted to certain demographics after application
			No
		Are simulations conducted prior to implementation to identify possible biases?	Yes, simulations designed for the specific use case
			Yes, but only general robustness simulations
			No
		Is there transparent documentation of the entire application processes?	Yes, review mechanisms and error sources are made public
			Yes, but made available only in summarized form
			Yes, but only for internal review
			No
		Are potential biases communicated?	Yes, publicly
			Yes, internally
			No
	Participatory procedures	Who has access to the AI application?	There is unrestricted access
			Access is restricted based on criteria correlated with protected categories (e.g. gender, race)
			Access is restricted based on protected categories
		Can anyone initiate an assessment of bias and processing of a complaint?	Yes
			Yes, but proof of being personally affected has to be provided
			No
		Is there a participation mechanism in place to include affected demographics?	Yes, on a regular basis
			Yes, but only once (e.g. at the beginning)
			No
		Are the stakeholders and affected demographics reliably defined?	Yes, there is a stakeholder documentation available for trusted intermediaries
			Yes, but no documentation
			No

2.1.3 Applying the VCIO approach to accountability as a value



ACCOUNTABILITY							Value
Corporate/institutional liability (retrospective)		Disclosure of internal organisational responsibilities (prospective)				Error tolerance	Criteria
Are appropriate monetary means, an insurance policy and/or other forms of compensation in place in case of liability?	Is there an ombudsperson?	Is there an institutionalised opportunity to provide anonymous information to relevant parties?	Are responsibilities defined with respect to third-parties (affected persons/users)?	Are responsibilities for possible damage and liability cases documented?	Is there a comprehensive logging of the design process?	Is there a culture of dealing openly with mistakes within organisations?	Indicators
Yes, sufficient financial resources are available	Yes, a respective body has been established and openly announced	Yes, a respective body has been established and openly announced	Yes	Yes	Yes, comprehensive logging of all incoming training and operating data, version control of data records, etc	Yes, errors can be addressed without excessive penalty threats	Observables
Yes, funds are available for typical or probable claims, but not for less probable scenarios	Yes, but access is only possible when fulfilling certain requirements	Yes, but access is only possible with difficulties or full security is not guaranteed	No, but there are other ways to contact responsible persons	No	Yes, logging/ version control from second parties (e.g. by data suppliers)	Yes, but openness to error leads to tolerance for error	
No	No	No	No, there is no office to contact		No, incoming data is not controlled or documented in any way	No, there is no sufficient focus on errors	

2.2 Values constituting the AI ethics rating

In the debate on ethical and trustworthy AI, various ethical principles and values have been put forward.[8] While these contributions all have their merits, for the rating, we specifically selected values that can be considered basic requirements for the prevention of harm and public welfare-oriented development of AI. Building on a meta-analysis of relevant publications, we settled on six values: transparency, accountability, privacy, justice, reliability, and environmental sustainability.

The following analysis of each value also takes into account further reaching problems concerning normative standards that regard the social effects of AI use. These include framework conditions, such as ensuring plurality of individual AI services, providers and infrastructures and questions of ecological and social sustainability in the entire AI development chain (raw material and energy consumption, click work in data farms), or the assessment of an AI system's resilience (reversibility of damage or more generally path dependencies). In the following discussion, we also capture crucial aspects of AI ethics in terms such as explainability, redress, or various tools to ensure privacy.

2.2.1 Transparency

Transparency to ensure participation and self-regulation

To realise the value of transparency is crucial for the fulfilment of societal goals like participation and self-regulation (Hustedt 2019). Transparency enables people affected by technical systems to adjust the AI's decision-making behaviour towards them in an enlightened manner, to identify and correct violations of rights, to engage in social debate or to build relationships of trust.

Explainability through debugging, simulability, decomposability and algorithmic transparency

Within the debate on ethical or trustworthy AI, transparency often combines demands for technical explainability of the algorithm itself, including requirements for the transparency of the development and training process of the AI. Explainability has risen to prominence in machine learning research because humans are so far not able to fully comprehend the technical "inner workings" of deep neural networks (DNN), also in unsupervised learning. As it is impossible to interpret exactly why deep neural networks produce individual results, various technical methods have recently been developed to mitigate this problem, converging under the said term of explainability (xAI).

Providing explainability aims to solve attribution problems, i.e. the ability to prove why particular errors have occurred (debugging), to increase confidence in technical systems or to be able to acquire new knowledge from methods of machine learning. Various other dimensions of explainability are being pursued (Mittelstadt et al. 2019) such as the mechanistic understanding of the functioning of machine learning models (simulability). Also, a model's components can be made comprehensible (decomposability) or learning algorithms be investigated (algorithmic transparency).

8 For a detailed analysis see Hagendorff 2020.

Regarding the development process, transparency concerns concrete questions: who is responsible for the development of machine learning models, who has approved the development, where does the data used to train models comes from, what quality tests have data sets undergone, who has labelled data sets, what learning objectives are pursued, what results are delivered by evaluations of models, what learning methods are used, how source code is viewed and much more (Pasquale 2015; Krafft and Zweig 2019; Burrell 2016; The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019, p. 247).

However, algorithmic processes cannot only be evaluated by checking source code, but also by "non-invasive" techniques. These include interviewing users, extracting data from accessible sources, systematic tests with "sock puppet accounts" used to examine certain aspects in isolation, among other methods (Lischka and Stöcker 2017, p. 55).

2.2.2 Accountability

Accountability as willingness or obligation to assume responsibility

In socio-technical environments, accountability does not necessarily refer to a causal responsibility of actors in the sense of agency. Instead, it typically means the willingness or obligation to assume responsibility (Saurwein 2018, p. 38; our translation). This entails specifying the obligations as well as designating the responsible agent. Especially in autonomous AI-based systems with unclear algorithmic selection and decision structures, it is crucial to counteract potential so-called responsibility diffusion and responsibility flight.

Accountable agents hold responsibility

From a prospective viewpoint, accountable persons or institutions have to be designated, who are then responsible for discharging the predefined transparency, verifiability and due diligence obligations associated with the AI system. Legal requirements, certification procedures, best practices, voluntary commitments, and other things may achieve this. As accountable agents, these bodies must also be identifiable to all those affected and easily accessible, i.e. for questions, complaints, or appeals.

Retrospective view includes liability and redress

From a retrospective viewpoint, questions of financial liability must be resolved, but also including considering non-monetary forms of redress (Floridi et al. 2018). Liability and redress mechanisms have to be designed independently of causal legal principles, such as negligence of action or fault in omission (e.g. AI insurance), as in some instances no clear culpability can be assigned. Here, concepts that hold institutions, companies or individuals liable for the operation or marketing of AI-based selection and algorithmic decision-making systems are necessary (risk/product liability).

Verifiability and replicability

Concerning current codes of ethics for AI applications, accountability includes aspects of adopting new regulations to an AI model's verifiability and replicability. These also include algorithmic impact assessments, the creation of monitoring bodies, the ability to appeal against algorithmic decision-making, or remedies for automated decisions (Fjeld et al. 2020).

2.2.3 Privacy

To ensure privacy AI ethics must consider several methods

Safeguarding an individual's private sphere is not only a necessary precondition for the protection of individual autonomy and agency, but also serves vital interests in self-development, self-realisation, engaging in intimate social relations as well as participate in democratic public deliberation.[9]

Informational privacy as data is being used for specific purposes, after explicit consent, and with a right to delete or rectify

For the question of ethical AI, the concept of informational privacy is of particular importance, since we often find AI applications in contexts of mass surveillance, healthcare, marketing and many other privacy-sensitive areas. At the same time, AI applications in themselves pose a privacy problem due to their reliance on large amounts of data. So to ensure privacy, AI ethics should consider several methods and principles: Personal data may only be collected and used for specific purposes. Once the purpose is fulfilled, the data may not be further processed. The data may only be used for a purpose other than that for which it was collected if the data subjects have given their explicit consent (specified and legitimate purpose). In addition to the principle of consent, privacy also includes the right to delete or rectify or the ability to restrict processing, meaning that individuals have the power to withhold their data from use in AI applications.

Furthermore learning with anonymous or pseudonymous data and using reliable anonymisation and pseudonymisation procedures should be promoted in the development of AI systems. In this regard, de-anonymisation potentials have to be taken into account.

Differential privacy and privacy by design

Privacy standards should be integrated into data processing itself, meaning privacy by design (Nissenbaum 2019). Additionally, research on using a mathematical definition of privacy that aims to maximise the usefulness of evaluating databases containing personal data, while ensuring that the individual records used cannot be identified is of great importance. The original data is obfuscated as much as necessary and then deleted; the evaluation is only done based on the obfuscated data and expressed as differential privacy.

2.2.4 Justice

Justice as algorithmic non-discrimination and question of fair working conditions

Questions of justice include problems of equal treatment and the fair distribution of certain goods. While existing AI ethics guidelines typically focus on questions of algorithmic non-discrimination and often frame those in terms of fairness, the VCIO approach broadens this perspective. This includes aspects of social justice, in particular, "hidden" work, which is essential for the operation of AI systems.

Discrimination in the negative sense then refers to distinctions and classifications that should not play a role in action because they are unjustified, i.e. they are based on stereotyping or degrading attributions, or based on attributes that shouldn't have a material impact on a decision. These are often linked to categories such as gender, age, ethnic or national origin, disability or pregnancy (Hagendorff 2019).

9 Rössler 2004, Arendt 2006, Fried 1984, Stahl 2016

Algorithmic discrimination through biases

Concerning discrimination by AI algorithms, reasons mostly lie in the reproduction of existing discrimination patterns that are introduced via the training data, in the (unintended) bias of software engineers, in the absorption of biases via presuppositions in labels, or the implementation of biases due to particular contexts of use. Other reasons relate to a lack of due diligence and thoughtfulness in the development process and a lack of completeness of data (bias in data selection). Such configuration flaws in software lead to unjustified different treatment of certain groups at the application level.

Click work as an aspect of social justice

The aspects of social justice that the VCIO model adds to this debate focus on the said "hidden" work that goes into the operation of AI systems. These services include precarious, potentially health-damaging labour, particularly so-called click work required for machine learning (Irani 2015). The training depends on the fact that large data sets are not only created but also annotated manually with labels (Engemann 2018). This is mostly done in specialised labelling companies, especially in Asia, which often exclude workers from minimum-wage or other workers' rights.

2.2.5 Reliability

Reliability as a precondition for trust

The consequences of erroneous outcomes, accidents or misuse of AI systems can affect individuals, parts of a system, or an entire society. Because of the different dimensions of harm, adequate strategies to build a reliable and trustworthy infrastructure have to be developed (Lampe and Kaminski 2019). As reliability is not only the precondition for trust in and/or predictability of the AI system but also a significant factor in the prevention of individual and societal harm, it is not only a technical but also an ethical principle.

Predictability and safety as robustness and resilience

AI applications are considered reliable when they perform in intended ways as well as when they do not possess vulnerabilities to external attackers. Reliability is akin to the concept of predictability, meaning that systems can prevent manipulation of various kinds. AI security problems arise when AI applications have software vulnerabilities, when they are not resilient against cyberattacks, or when the integrity and confidentiality of personal data are being compromised. AI applications, no different from any other intricate pieces of software, have security vulnerabilities. In most cases, we are talking about data poisoning attacks, adversarial examples or the exploitation of other flaws in the design of autonomous systems.

Safety deals with the preventability of accidents and unexpected system operations. Concerning AI systems, there are two aspects of safety: Robustness refers to the accuracy and reproducibility of the system's outcomes. Also, resilience gives a measurement of an AI system's error tolerance. This includes the ability of the technical backend to resist interferences, e.g. through component redundancy. A thorough technology assessment has to be implemented to reveal the most significant potential threats.

Cybersecurity as confidentiality, integrity and availability

The scientific community is working to counteract these threats, also in Cybersecurity. It is traditionally understood to include three aims concerning IT systems: confidentiality, integrity and availability. While confidentiality means that no unauthorised party has access

to the information, integrity covers aspects such as that information cannot be altered, that changes to the information are transparent and traceable, as well as the protection of the authenticity of the information. Availability, finally, refers to the accessibility of information and functionality when needed. These principles as optimisation targets can be compromised by failures, accidents or attacks from the in- and outside. Cybersecurity, therefore, relies on not only technical means, such as data encryption standards, firewalls, malware recognition, redundancy and disaster recovery measurements but also social practices and training.

2.2.6 Environmental sustainability

Resource-saving infrastructures to ensure intergenerational justice

Environmental sustainability is a form of intergenerational justice and describes the obligation towards future generations to ensure and preserve their living conditions. This obligation is typically geared towards a careful use of natural resources, e.g., to combat pollution and to preserve biodiversity as well as mitigate the worst effects of climate change.

Within the field of AI, this includes setting up resource-saving infrastructures for information technology, primarily through building power-efficient data centres as well as developing less power consuming machine learning models (Strubell et al. 2019). So far, the more computational resources AI models have at their disposal and the more training data they process, the more powerful and accurate the systems are. Increase in computation, however, means an increase in energy consumption, which brings with it increased carbon footprints. In this field, certification processes are especially useful for end-users to evaluate the carbon footprint of a given AI application. An important criterion to arrive at environment-friendly AI applications is the transparency regarding power consumption and the provision of sustainability data in general.

A right to repair

Another sustainability problem concerns the disposal of obsolete IT hardware, used to run AI applications (Crawford et al. 2018). In this context, a right to repair can improve the situation.

Positive effects of AI systems on the environment

While AI systems should not have a significant negative impact on sustainability and environmental protection goals, they can also be measured or valued by the extent to which they have positive effects on the environment. Computer vision can be used for tracking to detect illegal fishing vessels or bush fires via satellite imagery. Or it can be used for image and video classification, to identify endangered animals or poachers. Moreover, audio processing can be used to detect illegal logging. Those are just a few of many examples that show how various AI tools can be explicitly used to foster sustainability goals when taking the context into account.

2.3 How VCIO underpins the ratings in the AI Ethics Label

By discussing the VCIO model in more depth, we have set the foundation for a comprehensive approach towards handling AI ethics. Thereby, we demonstrated how these values might translate into practice through examples for specifying criteria and indicators. These can offer orientation to system developers and users alike and help regulators, watchdog and standard-setting organisations to specify their requirements for ethical and trustworthy algorithmic decision-making systems.

AI evaluation that citizens, users, and consumers understand

However, 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. We, therefore, propose capturing a core set of values in a standardised label, as shown previously. This is one among many possibilities how our VCIO model can be used for AI regulation. How to accurately conduct the rating and implement it in an AI ethics label is described in the following.

Levels of the ethics rating

The label includes one rating for each of the values captured with the VCIO-approach. Letters indicate each aggregated rating with "A" indicating (close to) complete value fulfilment. The label should include several levels, to sufficiently differentiate between different levels of value fulfilment and correspond with the granularity of the observables. But it should not include too many levels, as this might counteract its goal of providing an overview of a system's quality at a glance. Therefore, we suggest using a system with 5-7 levels (i.e. A to G).

How to aggregate the system rating

To define the levels of the system rating reached by a specific AI system, we need to aggregate the multiple observables subsumed under a particular value into a single rating to be displayed on the label. In principle, there are several ways in which to achieve this:

1) Different observables can have individual metric values, which we aggregate by computing the average value. As an example, school grades are aggregated in such a manner: A math test with a bad mark such as D can be offset by another test with a good mark such as a B, resulting in a final average math grade of C.

2) We can define the minimum requirements of observables needed to reach a specific system rating. Water quality ratings are aggregated in such a manner. A bad phosphate level results in a lower overall rating of water quality, independently from whether the salt level is high or low.

Minimum requirements aggregation as viable approach

►► ***Given these two alternatives, we recommend using a minimum requirements approach when aggregating observables into values. This is suitable as all indicators are technically equally important and necessary for an "ethical" AI system and their effects are interrelated. As necessary conditions, they cannot be set off against or substituted for each other.***[10]

10 As an example, if a system is checked for biases, but possible negative biases are not addressed, the effect would be the same as if biases had not been checked at all. Therefore, any level of rating needs a definition of minimum requirements to be met by the indicators to reach a certain rating of the value. However, indicators can be related to each other with regard to amplification, compensation or "killer" effects.

FIGURE 3 **The AI Ethics Label and the elements of the system rating**

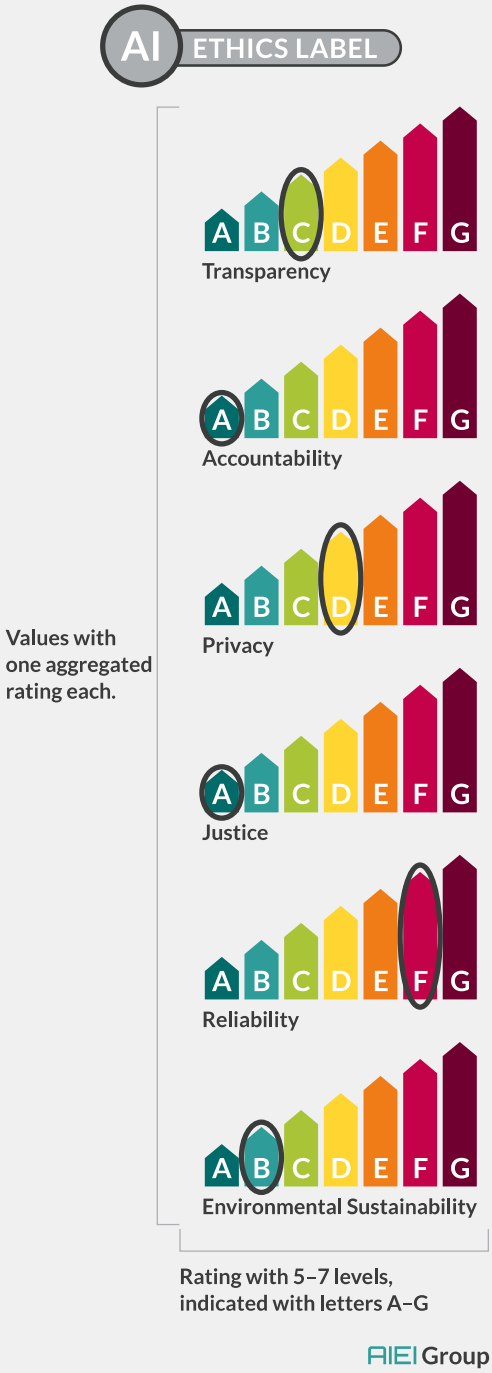
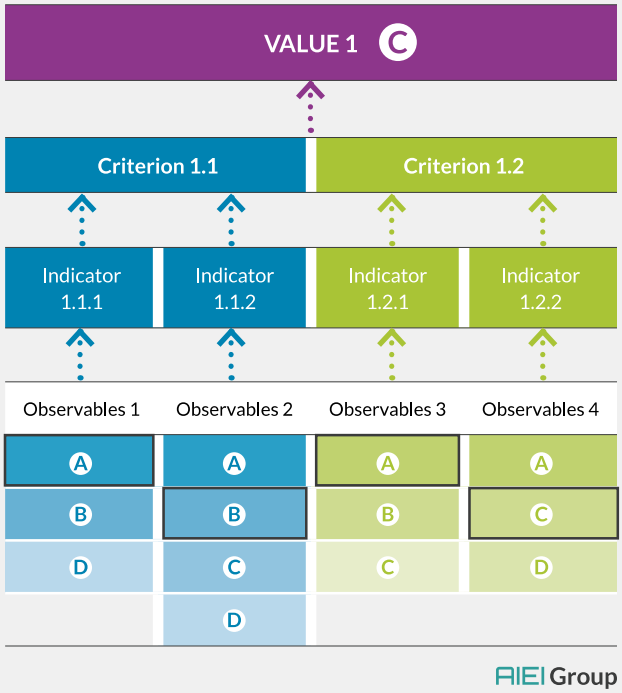


FIGURE 4 **System rating and operationalisation of a value using minimum requirements and aggregation**



Ensuring incentives for higher ratings despite aggregation

However, one drawback of a system using minimum requirements may be that it gives few incentives to strive for individual indicator ratings that go beyond the minimum requirements. For example, if a value has achieved an overall rating B, the AI system's provider has fewer incentives to strive for a better rating of a single indicator of said value as the label would not reflect this. However, different mechanisms could be considered to preserve incentives, such as indicating a certain number of higher-rated indicators with a '+' in the value's rating.

Still, to implement an AI ethics label, it is necessary to precisely set out the minimum requirements for each level of rating, thus specifying which level on the observable level leads to which level of value fulfilment.

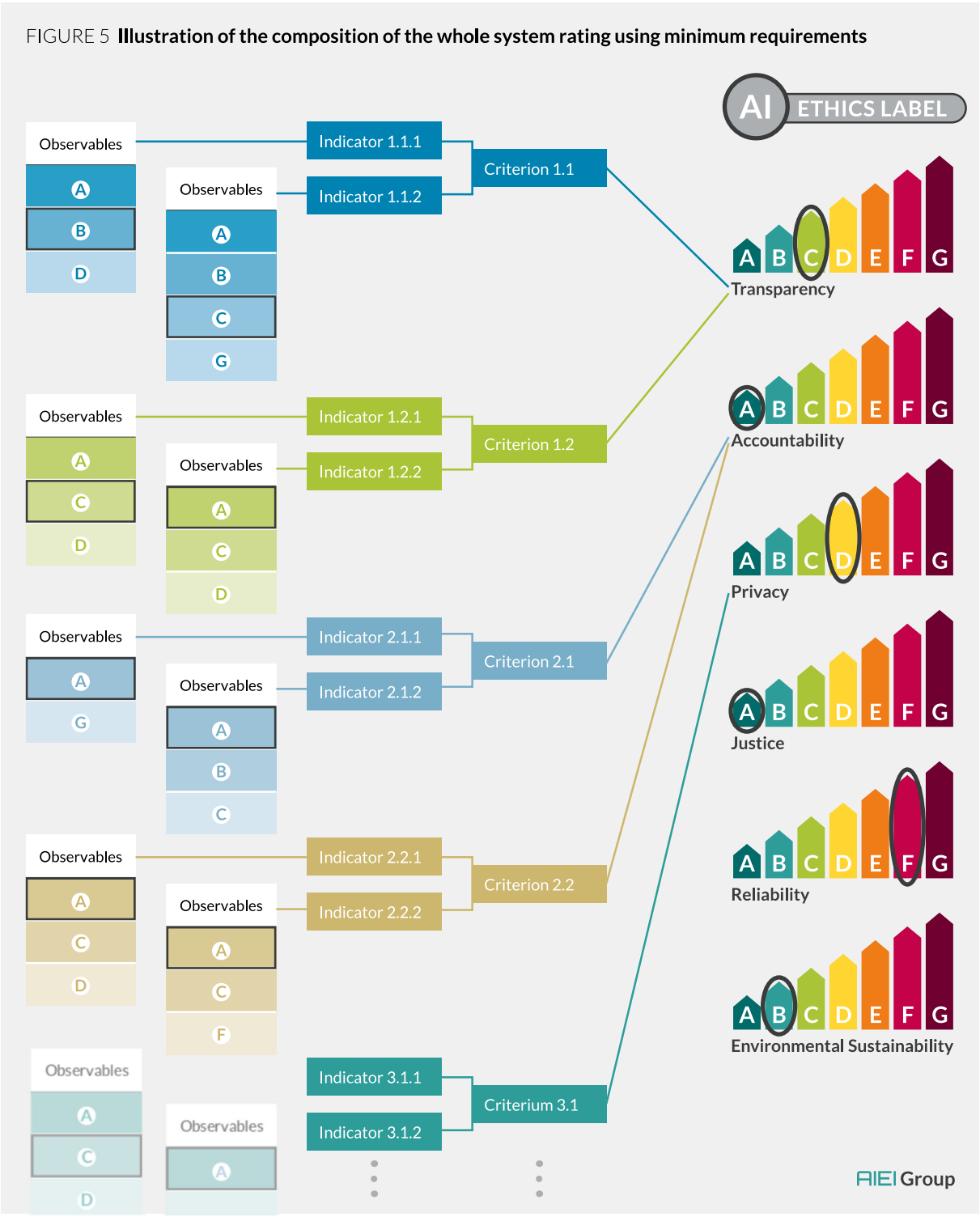
We can achieve this either through a bottom-up approach, whereby each observable is assigned a corresponding level on the value's rating, or a top-down approach. The latter requires to examine every level of the value rating and to define respective minimum requirements on the level of observables. Both approaches need an extensive and comprehensive negotiation process with all stakeholders involved.

An AI Ethics label can be given to an AI system independently from its future use. For example, a system used to determine the risk of relapse of convicted criminals receives the same rating whether it is being used in parole hearings and court cases or to determine reintegration measures after a convict is released from prison. We can, therefore, determine a rating before the AI system is being put to use.

AI system evaluation requires analysis of application context

However, the level required to ensure that the AI system is ethical much depends on the application context. For example, if an AI system is used for industrial processes it is subject to different requirements for transparency than if the same system used in medical procedures. In turn, the classification of the application context determines the level of value fulfilment needed for different applications. Only both steps together – the general description of the AI system with the Ethics Label and an assessment of the application context – can determine whether an AI system is ethical in a given situation. We describe the proposed methodology to classify the application context as captured in the risk matrix in the following chapter 3.

FIGURE 5 **Illustration of the composition of the whole system rating using minimum requirements**



3 CLASSIFYING AN AI'S APPLICATION CONTEXT

The AI Ethics Label provides at a glance information about the ethically relevant characteristics of an AI system. It displays a rating for each value, yet their relevance depends of where we apply the AI system. Take the scenario of an AI system in a specific medical context – it requires different levels of transparency than in some industrial applications. However, it is not feasible to consider these requirements for each application scenario, just as it is not feasible, for example, to write a criminal law that lists prison sentences for every conceivable case.

Risk measured as an AI system's potential harm (vulnerability) and affected persons' dependence on the decision (exposure)

Therefore, a viable approach for handling AI ethics requires an additional step: a classification of application contexts. This classification must be based on the overall potential damage an AI system may cause in its respective social process. Decisive factors in assessing this potential are the intensity of the potential harm of the AI system and the dependence of the affected person(s) on the respective decision. Here we found that using a two-dimensional risk matrix on which these factors describe the axes simplifies the classification process without abstracting too much from the given complexity an AI system operates in (Krafft and Zweig 2019).

Particularly for the legal certainty required by companies for the use of ethically uncritical AI systems, this classification must be sufficiently precise, but also quick to decide upon and implement. Considering the different fields of application and social contexts in which algorithmic decision-making systems can be used (e.g. advertising compared to medicine), it is essential that one solution may not fit all needs when it comes to governing the risks of ADM systems (Krafft and Zweig 2019, Saurwein et al. 2015, Van Drunen et al. 2019).

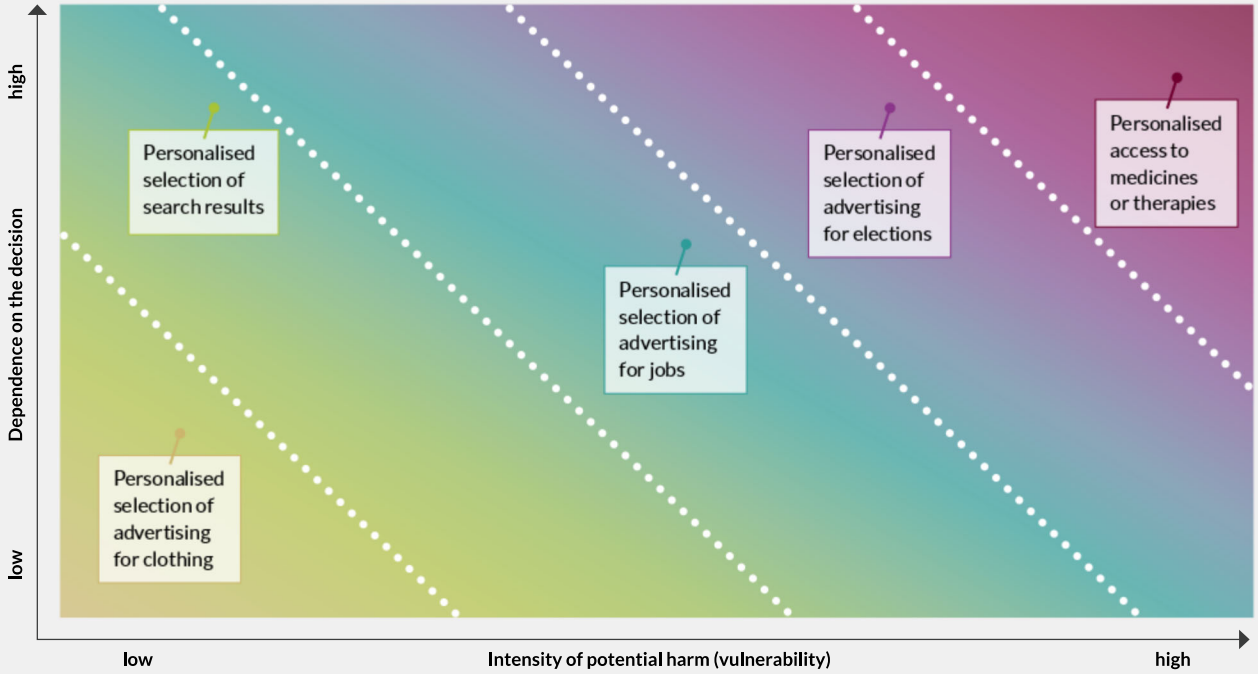
We define the details of the axes in section 3.2. Even though this project has not set itself the task of defining fixed categories, from our perspective, we can discern five classes as defined in our recommendations in section 3.3. These range from the exemption from a system rating for low-risk AI systems to application contexts where the use of algorithmic decision-making systems must be prohibited.

3.1 The risk matrix

Risk matrix to determine whether an AI system is ethically-sensitive within application context

The risk matrix serves to determine whether an AI system is ethically-sensitive with regard to its application context, i.e. it shows the risk for the potential damage of an ADM system depending on its usage. This classification also serves to determine whether and to what degree it requires regulation or labelling.

FIGURE 6 **Example of a risk matrix where the same technical component, i.e. recommendation algorithm, has different risk potentials in different areas of application**



Source: Krafft and Zweig 2019

AIEI Group

Using the example of recommendation systems, e.g. online searches and subsequent targeted advertising and recommendations, it quickly becomes clear how strong the influence of the usage scenario of ADM systems is.

Figure 6 shows ADM systems that could all be implemented based on the same IT components, i.e. the same recommendation algorithm. The only differences are the purpose of use, the training data used, and most likely also the system's quality requirements. So while there may be few severe ethical implications to consider when dealing with personalised suggestions for clothing, for example, these implications increase with the intensity of potential harm and a person's dependence on the decision. Take the use of recommendation systems for personalised suggestions for medical products, however, and it becomes clear that these systems need to be treated differently. So if for example, a regulatory body wants to use the proposed usage classes, these systems should be sorted into different "regulation" classes.

This risk matrix proposed by Krafft and Zweig (ibid. 2019) is an idealised scheme for defining classes of an ADM system according to their risk potential. The horizontal axis shows the risk depending on the intensity of potential harm. The vertical axis represents the dependence on the decision of persons affected (vulnerability).

3.2 Dimensions of the risk matrix

Risk matrix to reduce the complexity of application scenarios

Applying the risk matrix is an attempt to reduce the complexity of application scenarios for evaluation that involves considering multiple related aspects within each axis.

To handle underlying complexity, stakeholders must carefully weigh decisions with all affected parties

The two dimensions are naturally complex in their internal structure; correlations between the dimensions arise depending on the weight of individual aspects in the internal composition. This leads to a certain degree of vagueness in determining a hypothetically ideal dividing line between the degrees of classification. The process of determining such a line in practice requires the participation of stakeholders with a broad, interdisciplinary perspective.

In addition, the concept's complexity increases in cases of objective conflicts of values. Such conflicts affect both the intensity of potential harm and the dependence on the decision (x-axis and y-axis). They must be resolved in a way that takes the concerns of all parties involved and affected into account, and that preserves their pursued values as far as possible.

3.2.1 Intensity of potential harm (x-axis)

To assess the intensity of potential harm an AI system can have, look at the impact on a number of people or access to resources and whether society as a whole is threatened

For the x-axis, the critical aspect is potential harm, which regards the evaluation of the intensity with which an AI system could potentially harm people, organisations, and society. To assess this, the following issues must be regarded:

- Impact on fundamental rights, equality or social justice: Does an AI have a negative impact on a natural, legal persons' fundamental rights or are social justice mechanisms (e.g. pension, health insurance) at risk for extensive demographics or might the impact even be catastrophic and lead to loss of life (e.g. the treatment of intensive care patients)?

- Number of people affected: Is a high number of people affected (e.g. fair assessment for a job application)?

- Impact on society: Does the system bear the risk of affecting society as a whole (e.g. personalised selection of political news), independent of directly perceivable damage?

In any case, it is impossible to evaluate the intensity of potential harm by merely multiplying the amount of damage with the probability of occurrence. To do so would mean to equate the risk of someone leaving the house without an umbrella in case of an impending storm (high probability of occurrence, low potential damage) with the risk of a nuclear accident (low probability of occurrence, high potential damage). Consequently, as potential damage increases, macro risks can arise that threaten our ability to act at all and are, therefore, unacceptable.

3.2.2 Dependence on the decision (y-axis)

When assessing the dependence on the decision, look at control, switchability and redress

The y-axis shows the dependence of the potentially affected parties on the algorithmic decision, thus addressing the options to avoid the potential harm indicated on the x-axis. The better the chances are of avoiding exposure to the potential negative consequences of a decision or the damage caused by it, the further left on the y-axis the ADM system lands. The three main factors that play a role in assessing dependence on the decision are control, switchability and redress.

- Decisions and actions of an AI system additionally filtered through meaningful human interaction (e.g. the purchase of recommended items in an online shop) imply a lower demand for regulation than machines acting without human intermediaries (e.g. the emergency shutdown of a nuclear power station). This is expressed as control.

- The ability to change the AI system for another (e.g. by switching the operator) or avoid being exposed to an algorithmic decision altogether is called switchability. A one-sided relationship of dependence between producers or operators and users and monopolistic (including governmental) structures lead to dependence on one or a few systems. In the worst case, the user does not have the ability to opt-out of using specific services without facing societal repercussions (e.g. health care, financial market).

- The importance of the possibility of challenging or correcting an algorithmically-made decision and the time needed to follow up on the request adequately should not be underestimated and is known as redress. Machine-made decisions that cannot be challenged at all increase the dependence on the decision. To rectify significant individual harm takes more time and effort than many instances of lesser harm. This aspect concerns damage compensation/liability, as addressed in the dependence on the decision (y-axis).

3.3 Recommendation for classes

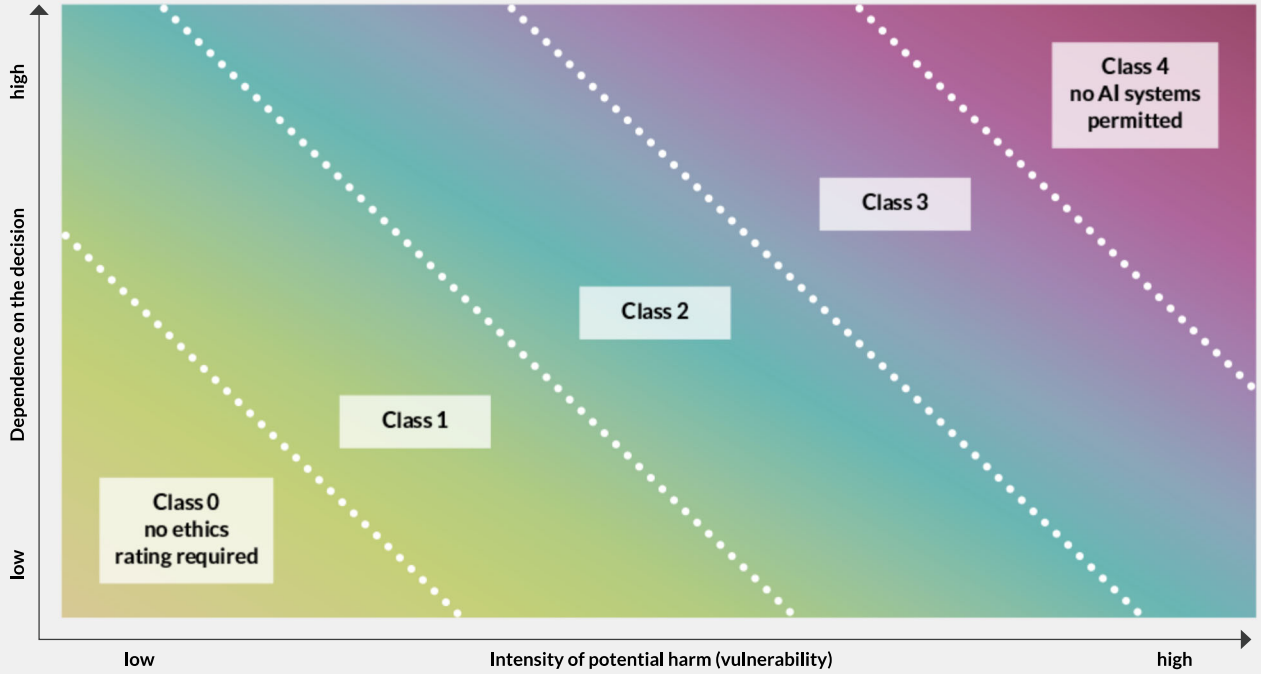
Five different risk classes emerge which require different degrees of regulation

Following the discussion of the axis, looking at the risk matrix as a whole, Krafft and Zweig see a division into five different classes emerging, as shown in Figure 7 (ibid. 2019).

Systems that do not require any regulation at all fall into class 0. The highest class, 4 in this case, serves to classify contexts where no AI system should be applied. For algorithmic decision-making systems that fall between these two extremes, a subdivision of at least three further classes seems to make sense to reflect increasing system requirements adequately.

Depending on the application scenarios, the minimum requirements for the fulfilment of different values (later displayed on the AI Ethics Label) can differ or can be regulated as cross-sectional specifications. For example, a condition could be that a system in class 2 must achieve at least a B in all ratings. Where necessary, the requirements for different values of the ethic rating, e.g. privacy and transparency, might also differ depending on the superordinate application scenarios (e.g. medicine or mobility).

FIGURE 7 **Risk matrix with 5 classes of application areas with risk potential ranging from 'no ethics rating required' in class 0 to the prohibition of AI systems in class 4**



Source: Krafft and Zweig 2019

AIEI Group

The decision, which kind of application falls into which class and is therefore subject to certain regulatory requirements, has to be made by regulators. We strongly recommend accompanying this process scientifically. In the following, we describe the classes in more detail.

Class 0 no ethics rating required

There is a multitude of AI systems for which the total damage potential, i.e. the intensity of potential harm and the exposure to the decision, is so low that a regulating instance can refrain from demanding a system rating.

▸▸ *In this case, we recommend not to require obligations for transparency, for example, or to install control processes permanently. In doubtful cases, a post hoc analysis should be carried out, and the risk assessment may have to be repeated. We expect most AI systems to fall into this class.*

Class 1

If the intensity of potential harm and the exposure to the decision exceed a certain threshold, a regulatory body should make initial demands for the ethics rating.

▸▸ *As an orientation, we recommend initial transparency obligations (Krafft and Zweig 2019), which include an interface for analysing the system as a black box and an explanation about how the ADM system is embedded in the social decision-making process. An AI system's embedded nature is also reflected in the VCIO model.*

Class 2

For the increasing intensity of potential harm and exposure to the decision, the input data must be fully disclosed (to the relevant audience as determined by the regulator), and the information regarding the ADM system's quality must be verifiable. If individual decisions can have a significant negative impact on individuals or groups, it is essential to ensure that an AI system is geared towards achieving objectives that minimise the damage.

▸▸ *As an orientation, we recommend transparency concerning the values applied by the AI system and the used training data set (Krafft and Zweig 2019). Also, there should be a possibility to review the quality assessment without having to rely on the values communicated by operators. They should provide the results of the algorithmic decision-making system in a form that enables the monitoring authorities to calculate, i.e. understand, the quality measures used as this clarifies the system's precise objectives.*

Class 3

In this class fall instances where either the potential (individual/societal) damage of decisions by the AI system is very high, the system is used without the knowledge of the persons affected, or where it can work against their expectations about the system. The aim must be to reduce risks as much as possible, i.e. to identify and avoid any way in which adverse decisions could be made.

▸▸ *As an orientation, we recommend monitoring and scrutinising the training and input data as well as the machine learning procedure (Krafft and Zweig 2019). Many machine learning methods cannot meet the transparency and explainability requirements necessary to allow for maximum risk avoidance. Only algorithmic decisions that are comprehensible and understandable by humans are permitted (Rudin 2019). All information must be comprehensible and verifiable at least for a panel of experts within a reasonable time frame. This requires various interfaces to the input data and the results of the machine-made decision.*

Class 4 no AI systems

Some ADM systems have such a high total damage potential that they should not be used with a machine learning component at all (e.g. autonomous weapon systems).

▸▸ *In these application contexts, regulators must prohibit the use of an algorithmic decision-making component. If a system is to be built that would fall into class 4, it needs to be adjusted in a way which reduces the intensity of potential harm and/or the dependence on the decision enough to justify categorisation into class 3.*

4 CONCLUSION AND WHERE TO GO FROM HERE

Our report has shown how to apply AI ethics in practice

This report demonstrates the transition from "what to how", as we focused on bringing abstract principles into technical and organisational practice.

We have shown how to apply AI ethics in a way which can help to:

- support the enforcement of European values and the protection of citizens in Europe
- create quality-transparency and comparability in the market
- does not impose an unnecessary burden on companies and is straightforward to implement where necessary
- is easy to communicate and understand.

The diagram summarises the three major elements of the model: characterising AI systems with an **ethics rating** based on the **VCIO** (values, criteria, observables, indicators) approach (left side), making the results easily understandable through the AI Ethics Label (circles on the AI Ethics label) and classifying the **application context** (right side), to determine the necessary rating requirements for a given AI system (blue lines on the AI Ethics Label).

4.1 Putting it all together

As previewed in section 1.3, this overall approach benefits a variety of different stakeholders:

- An **organisation planning to use an AI system** for a specific application follows an initial checklist to determine the ethical risk of the application context.If an application falls into the lowest risk level, the process ends at this stage. If there are ethical issues to consider, then the organisation performs a full assessment of the application context using the risk matrix.

- Similarly, **manufacturers of AI systems** can consider the range of expected applications using the risk matrix and decide whether to market an AI system only for applications without ethical sensitivity, or also for higher risk classes. In the latter case, they may gain market advantage by achieving a high ethics rating recognised worldwide for their products. This applies in both B2C and B2B settings.

- **Regulators** can use the combination of the risk matrix and ethics rating to specify requirements for different application contexts and to avoid over-regulation of application fields that do not pose any major ethical challenges. For application fields

classified in one of the higher risk levels, they may demand that an AI system must (1) carry an ethics label that shows the rating for values such as transparency, robustness, or fairness and (2) satisfy certain minimum levels within the rating.

- **Consumers** use the ethics label to compare AI products and services and make informed decisions about what is acceptable to them and/or worth spending money on. Consumers are alerted to ethically sensitive applications through the risk matrix. Moreover, consumers can trust that minimum regulatory requirements protect them.

- **Purchasers** (both in private and public sector procurement) use the ethics rating and risk matrix to create clear specifications and benefit from market transparency. As operators of AI systems, they have a recognised way of demonstrating ethical behaviour.

4.2 Next steps

Stakeholders need to come together

What needs to be done now? While we have developed the overall framework that can be used as a tool for organisations developing and using AI as well as policymakers, standard-setting bodies and other watchdog organisations, it is clear that several stakeholders need to cooperate. They need to refine and complete the approach and to put it into operation. In particular, we are aware that the VCIO examples given here are neither comprehensive nor detailed enough for direct application in the field. Instead, they would need to be further negotiated, specified and reflected.

We do not have all the answers to pressing questions, but we offer tangible tools and models to accelerate the discussion and to provide a foundation for the debate. Our model helps to measure values using criteria, indicators and observables and combines the context-based risk assessment so that nuanced AI ethics regulation is becoming an actionable course.

We see roles in particular for European standards developing organisations and European policymakers. As the authors of this report and coming together as AI Ethics Impact Group (AIEI Group), we see our future role in supporting standardisation and policy actors, initiating networks and activities, raising awareness, and refining the conceptual ideas.[11]

11 An issue that we have not addressed in detail so far is whether it is sufficient for an AI system manufacturer to claim an individual ethics rating, or whether an independent assessment is needed. Our initial hypothesis is a mix of both approaches, i.e. levels A to D would require independent assessment, while the manufacturer could simply declare levels E to G.

FIGURE 8 **Illustration of the composition of the whole system rating using minimum requirements**

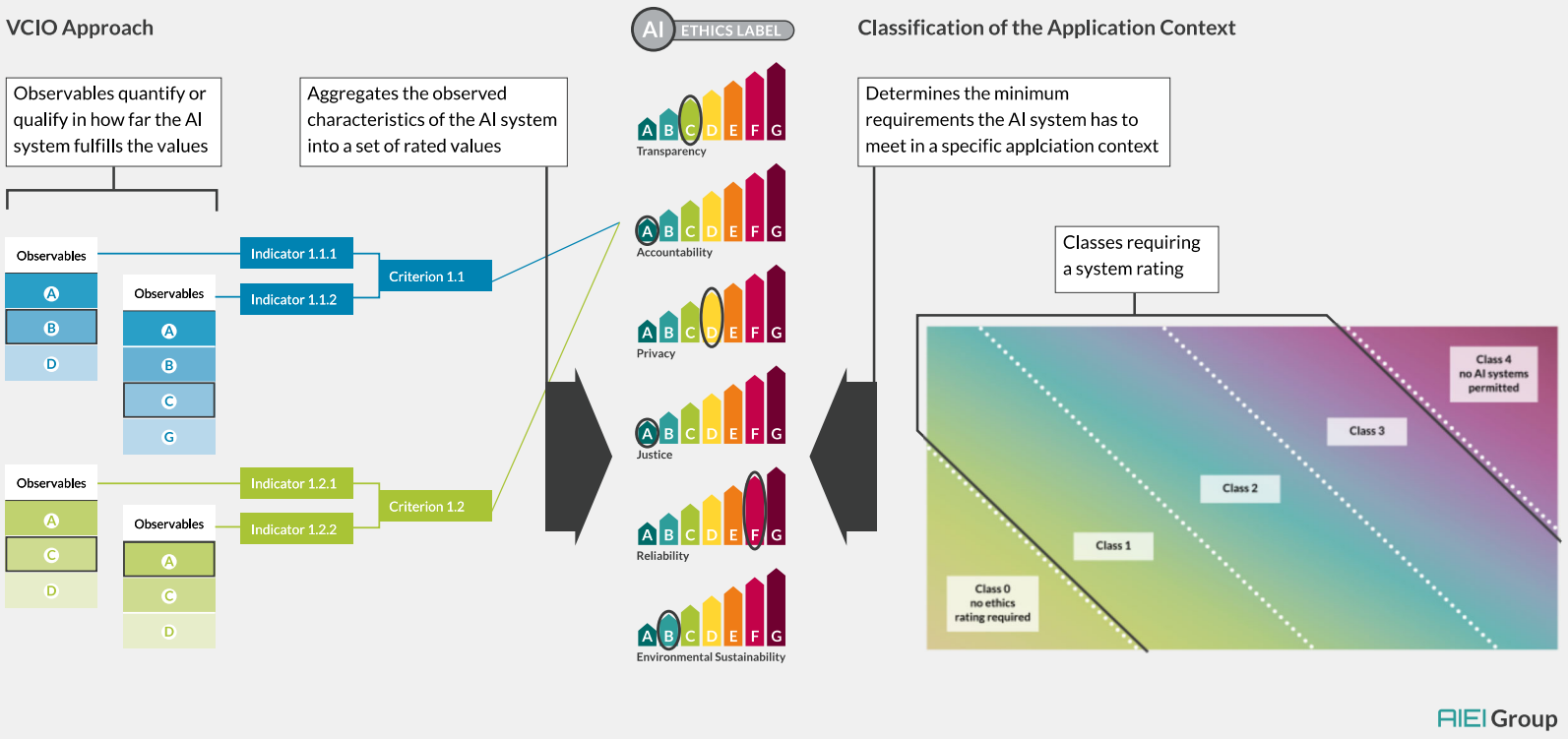


TABLE 1 **Specific Recommendations**

Stakeholder	Recommendations
Standards developing organisations	Industry and regulators have, for many decades, relied on standardisation. Established standard-setting bodies such as CEN, CENELEC and ETSI in Europe must bring together all relevant stakeholders. The aim is to reach and formulate consensus in such a clear and stringent way that it guides design and development and may provide a benchmark for testing and possibly certification. Well-established processes and structures ensure representation from industry, academia, and civil society and the outcome thus carries a certain weight and legitimacy. The European CEN-CENELEC AI Focus Group is currently preparing its initial report which covers some AI ethics issues from a standardisation perspective. Internationally, IEC Special Expert Group 10, as well as JTC1/SC42 WG3 and the IEEE P7000 committees, are working in this area, amongst others. Elements of the labelling and classification approach in this report have already been discussed and received broadly favourably in some of these standardisation committees. **We propose that European standardisation committees bring together relevant experts and stakeholders with a focus on refining and completing the VCIO-based descriptions of the values in the ethics rating. These committees could also discuss and further refine the selection and naming of values.** Given the urgency of addressing ethical challenges of AI, standardisation committees need to be well supported in this task.
Policymakers, especially the European Commission and European Parliament	The European Commission already mentions labelling as an element for AI ethics in the February 2020 version of its AI white paper. However, this initial version could be interpreted as opening the doors for simple kitemark schemes which would not be adequate. **We propose that the next version of the white paper clarifies the approach to AI ethics labelling, drawing on the framework we have presented in this report.** The bimodal high-risk/low-risk classification of AI applications in the initial version of the white paper appears as an oversimplification. **We propose a classification of application contexts with 4 or 5 levels based on a small number of horizontal criteria as outlined above. This makes a sectoral approach mostly unnecessary.** It has been good practice for many years to have a "division of labour" between standardisation and regulation. While standardisation deals with the full depth of technical issues and builds consensus among experts, the regulation gives "teeth" to standards by referring to them without having to specify technical details. **We, therefore, propose that the European Commission supports European standardisation committees in their work on refining and completing this framework for AI ethics and at the same time prepares legislation to give it a similar status as, for example, the energy efficiency label.**

5 BIBLIOGRAPHY

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D. (2017) Concrete Problems in AI Safety [Online]. Available at: https://arxiv.org/pdf/1606.06565.pdf (Accessed: 03 March 2020).

Arendt, H. (2006) 'The Crisis in Education' in Arendt, H. (ed.) Between past and future. Eight exercises in political thought. With assistance of Jerome Kohn. 10th Edition. New York: Penguin Books, p. 170–193.

Beijing Academy of Artificial Intelligence (BAAI) (2019) Beijing AI Principles [Online]. Available at: https://www.baai.ac.cn/blog/beijing-ai-principles (Accessed: 03 March 2020).

Burrell, J. (2016): How the machine 'thinks'. Understanding opacity in machine learning algorithms. In Big Data & Society 3 (1), pp. 1–12.

Crawford, K., Joler, V. (2018) Anatomy of an AI System. AI Now [Online]. Available at: https://anatomyof.ai/ (Accessed: 03 March 2020).

Engemann, C. (2018) 'Rekursionen über Körper. Machine Learning Trainingsdatensätze als Arbeit am Index' in Engemann, C., Sudmann, A. (eds.) Machine Learning – Medien, Infrastrukturen und Technologien der Künstlichen Intelligenz. Bielefeld: Transcript, p. 247–268.

European Commission (2019) Ethics Guidelines for Trustworthy AI. Luxembourg: High Level Expert Group on Artificial Intelligence (AI HLEG). [Online]. Available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation (Accessed: 03 March 2020).

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M. (2020) 'Principled Artificial Intelligence. Mapping Consensus in Ethical and Rights-Based Approaches to Principles for Ai' in Berkman Klein Center Research Publication 2020(1), p. 1–39.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V. et al. (2018) 'AI4People – An Ethical Framework for a Good AI Society. Opportunities, Risks, Principles, and Recommendations', Minds and Machines 28(4), p. 689–707.

Fraunhofer IAIS (2019) Vertrauenswürdiger Einsatz von Künstlicher Intelligenz (white paper) [Online]. Available at: https://www.iais.fraunhofer.de/content/dam/iais/KINRW/Whitepaper_KI-Zertifizierung.pdf (Accessed: 06 March 2020).

Fried, C. (1984) 'Privacy. A moral analysis' in Schoeman, F. (ed.) Philosophical dimensions of privacy. An anthology. Cambridge: Cambridge University Press, p. 203–222.

Friedman, B., Nissenbaum, H. (1996) 'Bias in computer systems', ACM Transactions on Information Systems, 14(3), p. 330–347.

Graham, M., Hjorth, I., Lehdonvirta, V. (2017) 'Digital labour and development: impacts of global digital labour platforms and the gig economy on worker livelihoods', Transfer: European Review of Labour and Research 23 (2), p. 135–162.

Hagendorff, T. (2019) 'Maschinelles Lernen und Diskriminierung: Probleme und Lösungsansätze', Österreichische Zeitschrift für Soziologie, 44(1), p. 53–66.

Hagendorff, T. (2020) 'The Ethics of AI Ethics. An Evaluation of Guidelines', Minds and Machines, p. 1–22. [Online]. Available at: https://arxiv.org/pdf/1903.03425.pdf (Accessed: 03 March 2020).

Heesen, J. (2012) 'Computer and Information Ethics' in Chadwick, R. (ed.) Encyclopedia of Applied Ethics. 2nd Edition. San Diego: Academic Press, p. 538–546.

Hubig, Ch. (2007) Die Kunst des Möglichen Bd. 2., Ethik der Technik als provisorische Moral. Bielefeld: transcript-Verlag.

Hubig, Ch. (2016) 'Indikatorenpolitik', CSSA Discussion Paper 2016(2), [Online]. Available at: https://www.cssa-wiesbaden.de/fileadmin/Bilder/B%C3%BCcher_Brosch%C3%BCren/Papers-cssa/cssa-paper_Indikatorenpolitik_2_2016.pdf (Accessed: 06 March 2020).

Hustedt, C. (2019) Algorithmen-Transparenz. Was steckt hinter dem Buzzword? [Online]. Available at: https://algorithmenethik.de/2019/05/06/algorithmen-transparenz-was-steckt-hinter-dem-buzzword/ (Accessed: 09 March 2020).

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) Ethically Aligned Design [Online]. Available at: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html (Accessed: 03 March 2020).

Irani, L. (2015) 'The cultural work of microwork'; New Media & Society, 17(5), p. 720–739.

Jobin, A., Ienca, M., Vayena, E. (2019) 'The global landscape of AI ethics guidelines', Nature Machine Intelligence 1(9), p. 389–399.

Kaminski, A. (2019) 'Begriffe in Modellen. Die Modellierung von Vertrauen in Computersimulation und maschinellem Lernen im Spiegel der Theoriegeschichte von Vertrauen' in Saam, N. J., Resch, M., Kaminski, A. (Eds.) Simulieren und Entscheiden. Entscheidungsmodellierung, Modellierungsentscheidungen, Entscheidungsunterstützung. Wiesbaden: Springer, p. 167–192.

Krafft, T. D., Zweig, K.A. (2019) , Transparenz und Nachvollziehbarkeit algorithmen basierter Entscheidungsprozesse | Ein Regulierungsvorschlag (vzbv) [Online]. Available at: https://www.vzbv.de/sites/default/files/downloads/2019/05/02/19-01-22_zweig_krafft_transparenz_adm-neu.pdf (Accessed: 03 March 2020).

Lampe, H., Kaminski, A. (2019) 'Verlässlichkeit und Vertrauenswürdigkeit von Computer-simulationen'. In Liggieri, K., Müller, O. (Eds.): Handbuch Mensch-Maschine Interaktion. Stuttgart, Weimar: Metzler p. 325–331.

Lischka, K., Stöcker, C. (2017) Digitale Öffentlichkeit. Wie algorithmische Prozesse den gesellschaftlichen Diskurs beeinflussen (Arbeitspapier). Gütersloh: Bertelsmann Stiftung, pp. 1–88.

Nissenbaum, H. (2019) 'Contextual Integrity Up and Down the Data Food Chain', Theoretical Inquiries in Law 20(1), p. 221–256.

Mittelstadt, B., Russell, C., Wachter, S. (2019) 'Explaining Explanations in AI' Proceedings of the Conference on Fairness, Accountability, and Transparency – FAT*, 19, pp. 1–10.

OECD (2019) OECD Council Recommendation on Artificial Intelligence [Online]. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (Accessed: 03 March 2020).

Pasquale, F. (2015) The Black Box Society. The Secret Algorithms That Control Money and Information. Cambridge, Massachusetts: Harvard University Press.

Rössler, B. (2004): The value of privacy. 1st edition. Cambridge: Polity Press.

Rudin, C. (2019) 'stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', Nature Machine Intelligence 1, p. 206–215.

Saurwein, F., Just, N. and Latzer, M. (2015) 'Governance of algorithms: options and limitations', info, 17(6), p. 35-49.

Stahl, T. (2016) 'Indiscriminate mass surveillance and the public sphere', Ethics and Information Technology 18(1), p. 33–39.

Strubell, E., Ganesh, A., McCallum, A. (2019) Energy and Policy Considerations for Deep Learning in NLP [Online]. Available at: https://arxiv.org/pdf/1906.02243.pdf (Accessed: 03 March 2020).

Van de Poel, I. (2013): 'Translating Values into Design Requirements' in Diane, P. and Michelfelder (ed.) Philosophy and engineering: reflections on practice, principles and process. 15th Edition. Dordrecht: Springer, p. 253–266.

Van Drunen, M. Z., Helberger, N., Bastian, M. (2019) 'Know your algorithm: What media organisations need to explain to their users about news personalization', International Data Privacy Law, p. 1–16.

6 ABOUT THE AUTHORS

Algorithm Accountability Lab at TU Kaiserslautern – Tobias Krafft and Marc Hauer

The Algorithm Accountability Lab under the supervision of Prof. Katharina Anna Zweig at the TU Kaiserslautern joins researchers for algorithmic decision-making systems, artificial intelligence, and neural networks. The institute was in a prime position to contribute to this report, primarily in the form of the risk matrix, thus adding context to any AI application and regulation according to the VCIO model. It builds on previous research by Prof. Zweig and Krafft, which included a regulatory proposal for AI systems from a socio-informatics (social informatics) perspective for the Federation of German Consumer Organisations (vzbv).

Tobias Krafft is among the first trained "socio-informaticians" (social computer scientists) in Germany. He received the Weizenbaum Study Award 2017 for his master thesis on the poor quality of algorithmic prediction systems in US courts. He is currently working on his doctorate concerning algorithmic accountability in various application contexts.

Marc Hauer is a doctoral candidate at the Algorithm Accountability Lab focusing on developing concepts for software development processes which help to exclude problems, non-technical errors or, unintended side effects in algorithmic decision-making systems that exceed moral or even legal limits as reliably as possible.

Bertelsmann Stiftung | Ethics of Algorithms – Carla Hustedt and Lajla Fetic

With its 'Ethics of Algorithms' project, the Bertelsmann Stiftung is examining the societal consequences of algorithmic decision-making to ensure the use of these systems serves society. The research involves work to help inform and advance algorithmic systems that facilitate greater social inclusion and committing to what is best for society rather than what is technically possible.

Carla Hustedt leads the Ethics of Algorithms project of the Bertelsmann Stiftung, Europe's largest operational think tank. Together with partners from diverse disciplines and sectors she and her team are conducting research on the societal consequences of algorithmic decision-making while also working on practical solutions for putting machines in the service of humankind. In 2019, Carla coordinated the development of the Algo.Rules, a set of nine principles for the ethical development of algorithmic decision-making systems. She also consulted the German parliament's AI Enquete Commission on the issue of AI-Transparency. Carla holds a Master in Public Administration from the London School of

Economics and a Master in Public Policy from the Hertie School of Governance. In 2019 Carla was a finalist in the category 'Science' at the Digital Female Leader Award and has been named the German representative on the 2020 list "100 brilliant women in AI Ethics" by Lighthouse.

Joining Hustedt as project manager was **Lajla Fetic**, who has previously worked on the topics of digitalisation and automation. She leads the development of tools for the ethical implementation of algorithmic systems in the public sector. Previously she worked for an international public sector consultancy on the digital transformation of the industry sector and for a think tank at the intersection of technology and society. Lajla is currently completing a Professional Year at the Bertelsmann Stiftung as part of her Master in Public Policy at the Hertie School of Governance with a focus on e-government and public sector innovation.

TU Darmstadt – Professor Emeritus Christoph Hubig

Professor Christoph Hubig is Professor Emeritus of Practical Philosophy/Philosophy of Scientific Culture at the TU Darmstadt. He continues to head the commission 'Revision of the ethical principles of the engineering profession/inclusion of AI and the design of autonomous systems' at the VDI association. The procedure proposed in this report is consistent with the VDI's approach to develop and offer a model value comparison based on "self-orientation", i.e. not to anticipate or prescribe solutions through specifications.

Professor Hubig is also a member of the commission 'Philosophy of digitisation and artificial intelligence' at CAIS Bochum and also partners with the Ethics Centre (IZEW) of the University of Tübingen in the area 'Ethics of AI/machine learning'.

High-Performance Computing Center Stuttgart (HLRS) – Dr Andreas Kaminski and Michael Herrmann

The High-Performance Computing Center Stuttgart (HLRS) of the University of Stuttgart provides high-performance computing platforms and technologies, services and support to researchers across Europe. The HLRS is among the most advanced research, development and service facilities in Germany in the field of simulation, visualisation and data analytics applied in the health, environment, energy, and mobility sectors.

To investigate social, political and philosophical aspects of emerging technologies such as AI, the HLRS established an in-house department for the philosophy of science and technology, which is headed by **Dr Andreas Kaminski**. Together with HLRS director, **Prof. Dr Michael Resch**, he is investigating the changes in science and society triggered by computer-intensive methods. Since 2010 he has been publishing on the shift in human-machine interaction through learning algorithms.

Further contributions were made by **Michael Herrmann**, a doctoral candidate in the department Philosophy of Science & Computer Simulation at HLRS. He graduated in mathematics and philosophy. In his PhD, he is studying the intimate relationship between

mathematics and technology within computer simulations and machine learning methods. He is also a lecturer in Engineering science at the University of Stuttgart.

International Center for Ethics in the Sciences and Humanities (IZEW) – PD Dr Jessica Heesen, Dr Thilo Hagendorff and Dr Wulf Loh

The International Center for Ethics in the Sciences and Humanities, Ethics Center for short, is an interdisciplinary research centre at the Eberhard Karls University of Tübingen. Research at the facility concerns ethical questions on science and its effects and builds on experiences from many years in the field. The Uni Tübingen is part of the Cluster of Excellence Machine Learning.

PD Dr Jessica Heesen teaches at the IZEW and acts as Head of Media Ethics and Information Technology at the Ethics Centre. She has directed several research projects on value-oriented development of AI and has contributed her findings to this paper. She is also a member of the 'Forum Privacy' of the German Federal Ministry of Education and Research.

Joining Dr Heesen as a contributor was **Dr Thilo Hagendorff**, who works as a technology ethicist for the 'Machine Learning: New Perspectives for Science' Excellence Cluster at the University of Tübingen and is also part of the Ethics Centre staff.

Further contributions were made by **Dr Wulf Loh**, a PostDoc researcher at the IZEW and involved in several AI and projects in human-robot interaction (HRI). Among his research interests are media and AI ethics, especially concerning values such as privacy, democratic participation, and discrimination.

iRights.Lab – Philipp Otto and Michael Puntschuh

The iRights.Lab is an independent think tank that conducts applied research to develop strategies and solutions for positively shaping changes in the digital world. It supports public institutions, foundations, enterprises, research institutions, and policymakers by bringing together legal, technical, economic and socio-political perspectives on digital issues.

The iRights.Lab founder and director, **Philipp Otto**, works on strategic individual concepts and models for dealing with the challenges digitalisation poses. He was a Visiting Researcher at the Berkman Centre for Internet & Society at Harvard University and has since published books and papers on political strategies, including the essay collection '3TH1CS – A reinvention of ethics in the digital age'.

Michael Puntschuh is a policy analyst at iRights.Lab. His research includes human rights in cyberspace or legal and ethical issues of digital technologies (governance). He was involved in the development of Algo.Rules, design criteria for algorithmic systems, and contributed to the issues surrounding the application of the VCIO-approach to values in this report.

Institute of Technology Assessment and Systems Analysis (ITAS) – Professor Rafaela Hillerbrand, Paul Grünke and Torsten Fleischer

The Institute of Technology Assessment and Systems Analysis (ITAS) at the KIT is a leading research institute looking at scientific and technical developments that concern systemic interrelations and technology impacts. Research at ITAS focuses on ethical, ecological, economic, social, political-institutional, and cultural issues and the institute aims to provide advice on research and technology policy, on the design of socio-technical systems, and to conduct discursive procedures on open or controversial technology policy issues. ITAS has run the Office of Technology Assessment for the German parliament (Bundestag) for many years.

Professor Rafaela Hillerbrand holds a PhD in theoretical physics and a PhD in philosophy and also held a position as a senior research fellow at the University of Oxford. Now at ITAS and as part of the AI Ethics Impact Group, she contributes years of advising politics and industry on the responsible handling of technology, especially in the context of digitisation.

Paul Grünke is a doctoral candidate in Prof. Hillerbrand's group at the Karlsruhe Institute of Technology. He has a background in mathematics and philosophy of science and works on the epistemology of computer simulations and machine learning in his PhD.

Torsten Fleischer heads the research area 'Innovation Processes and Technology Impacts' at ITAS, where he is primarily concerned with technology impact assessments for the computerisation and automation of mobility and transport. He also looks at interactions between technical and social change and the governance of innovation processes, including AI ethics.

VDE Association for Electrical, Electronic & Information Technologies (VDE e. V.) – Dr Sebastian Hallensleben and Andreas Hauschke

Counting 36,000 members, the VDE is one of the largest technical and scientific associations in Europe. It combines science, standardisation, and product testing with the main topics ranging from the energy transition over Industry 4.0 and smart technologies as well as artificial intelligence and digitisation. The technology association acts as the German member of IEC, CENELEC, and ETSI and is thus responsible for significant harmonisation efforts on an international scale.

At VDE, **Dr Sebastian Hallensleben** oversees the artificial intelligence portfolio and as such is the convenor of the AI Focus Group at European level through CEN-CENELEC. He further convenes the IEC SEG 10 efforts for the international standardisation of AI ethics and thus supports the labelling approach for this paper. He has been leading the project that has resulted in this report.

Also at VDE, **Andreas Hauschke** is concerned with the explainability of artificial intelligence and the requirements for the regulation and testing of AI systems. He graduated as an industrial engineer with a thesis on the interpretation methods of convolutional neural networks.

Joining the project in her capacity as an editor was **Nora Manthey**, who contributed a fresh look at the Group's work to build bridges for the readers. As a political scientist and producer in creative technology innovation (MFA), issues of human-machine interaction profoundly affect her practice.

Imprint

Bertelsmann Stiftung 2020
Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Phone +49 5241 81-0
www.bertelsmann-stiftung.de

Responsible
Dr Sebastian Hallensleben (VDE e. V.)
Carla Hustedt (Bertelsmann Stiftung)

Authors
Dr Sebastian Hallensleben
Carla Hustedt

Lajla Fetic
Torsten Fleischer
Paul Grünke
Dr Thilo Hagendorff
Marc Hauer
Andreas Hauschke
PD Dr Jessica Heesen
Michael Herrmann
Prof. Dr Rafaela Hillerbrand
Prof. Emeritus Christoph Hubig
Dr Andreas Kaminski
Tobias Krafft
Dr Wulf Loh
Philipp Otto
Michael Puntschuh

Editing
Nora Manthey

Graficdesign
Nicole Meyerholz, Bielefeld

Address | Contact

Dr Sebastian Hallensleben
Head of Digitalisation and AI
VDE Association for Electrical,
Electronic & Information Technologies e. V.
Stresemannallee 15
60596 Frankfurt am Main
Germany
Phone +49 69 6308 305
Mobile +49 170 7916306
sebastian.hallensleben@vde.com

Carla Hustedt
Project Lead Ethics of Algorithms
Programm Megatrends
Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Germany
Phone +49 5241 81-81156
carla.hustedt@bertelsmann-stiftung.de