Fraunhofer
IAIS

FRAUNHOFER INSTITUTE FOR INTELLIGENT ANALYSIS AND INFORMATION SYSTEMS IAIS

TRUSTWORTHY USE OF ARTIFICIAL INTELLIGENCE

PRIORITIES FROM A PHILOSOPHICAL, ETHICAL, LEGAL, AND TECHNOLOGICAL VIEWPOINT AS A BASIS FOR CERTIFICATION OF ARTIFICIAL INTELLIGENCE

In cooperation with

UNIVERSITÄT BONN

Universität zu Köln

Powered by

KI NRW

Sponsored by

Ministerium für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen

Fraunhofer IAIS

FRAUNHOFER INSTITUTE FOR INTELLIGENT ANALYSIS AND INFORMATION SYSTEMS IAIS

TRUSTWORTHY USE OF ARTIFICIAL INTELLIGENCE

PRIORITIES FROM A PHILOSOPHICAL, ETHICAL, LEGAL, AND TECHNOLOGICAL VIEWPOINT AS A BASIS FOR CERTIFICATION OF ARTIFICIAL INTELLIGENCE

Authors

Prof. Dr. Armin B. Cremers | Fraunhofer IAIS
Dr. Alex Englander | University of Bonn
Prof. Dr. Markus Gabriel | University of Bonn
Dr. Dirk Hecker | Fraunhofer IAIS
PD Dr. Michael Mock | Fraunhofer IAIS
Dr. Maximilian Poretschkin (project leader) | Fraunhofer IAIS
Julia Rosenzweig | Fraunhofer IAIS
Prof. Dr. Dr. Frauke Rostalski (project leader) | University of Cologne
Joachim Sicking | Fraunhofer IAIS
Dr. Julia Volmer | Fraunhofer IAIS
Jan Voosholz (project leader) | University of Bonn
Dr. Angelika Voss | Fraunhofer IAIS
Prof. Dr. Stefan Wrobel | Fraunhofer IAIS

In cooperation with

UNIVERSITÄT BONN

Universität zu Köln

Powered by

KI NRW

Sponsored by

Ministerium für Wirtschaft, Innovation, Digitalisierung und Energie des Landes Nordrhein-Westfalen

CONTENTS

WELCOME

Dear readers,

We live in the age of digitalization. Data are the fuel and artificial intelligence (AI) is the engine when it comes to using the resource of data for the benefit of the economy and society. All studies agree that artificial intelligence will increase global economic growth to a significant extent. Furthermore, it has the potential to help us tackle the major societal challenges like climate change, mobility, and health. While we need to unlock this potential, it is important to put people at the center of the developments. We need social dialog about the relationship of humans and machines along with a reliable ethical/legal foundation for this major future issue.

Today, North Rhine-Westphalia is already among the leaders in the development and application of artificial intelligence. In our state, we have first-class colleges and non-university research institutes where cutting-edge AI research is carried out with international visibility. In the economy, several large companies from North Rhine-Westphalia have built up extensive expertise in the field of artificial intelligence and have followed the path of digital transformation. Also, the framework in North Rhine-Westphalia is highly suited for start-ups and many new business models based on artificial intelligence have established themselves here. Using the KI.NRW competence platform that we have initiated, we are networking players in the artificial intelligence field and boosting technology transfer from research to practice as well as professional qualification. Nevertheless, artificial intelligence will only increase our prosperity and quality of life if values like self-determination, freedom from discrimination, data protection, safety, and security are taken into account.

Through the certification system for artificial intelligence that we have initiated, we want to further establish the quality mark "AI made in Germany" from here in North Rhine-Westphalia by identifying reliable, safe technology, and protecting it in a sustainable way. The certification system will support free competition among different suppliers and will make a contribution to the acceptance of artificial intelligence in society.

The development of the certification system was masterminded by experts from the fields of machine learning, law, philosophy, ethics, and IT security. The basic principles for technically reliable and ethically responsible artificial intelligence will be developed in an openly organized process that involves a wide range of stakeholders from business, research, and society. We are very pleased to launch and sponsor this initiative, which has an appeal reaching beyond Germany, from here in North Rhine-Westphalia.

This publication forms the basis for development of the AI certification system. It explains the audit areas to be dealt with in achieving trustworthy use of artificial intelligence. At the same time, I would also like to encourage you to take part in the social discourse on this future technology that we in North Rhine-Westphalia want to shape together with you through dialog.

Sincerely,

Prof. Dr. Andreas Pinkwart
Minister of Economic Affairs, Innovation, Digitalization and Energy of the State of North Rhine-Westphalia

FOREWORD

Dear readers,

Artificial intelligence (AI) is fundamentally changing society, the economy, and everyday life. It is also creating great opportunities for the way we live and work together. For example, it helps doctors evaluate X-rays better and often more accurately. It answers questions on insurance policies and other products by means of chatbots. Also, in the foreseeable future, it will enable cars to become more and more autonomous. At the same time, it is becoming increasingly clear that these applications need to be designed carefully so that we can make use of the opportunities AI brings while still respecting our social values and views.

Artificial intelligence has the potential to extend human capabilities and help us make new discoveries. Making decisions based on its results, which are automated or semi-automated through machine learning, also sets fundamentally new challenges for us. In addition to questions of technical suitability, general philosophical/ethical considerations come to the fore as well as legal issues. The possibility that intelligent machines will react "autonomously" casts new light on the individual liability and responsibility of people and thus on the basis and criteria of "attribution". To ensure that humans are always at the center of this development, we therefore need close communication about artificial intelligence between the areas of IT, philosophy, and law.

In view of the fast advance of artificial intelligence in almost every area of society, we have set the goal of developing certification for artificial intelligence in interdisciplinary exchange. This publication forms the beginning of this and looks at current challenges for artificial intelligence from the viewpoint of IT, philosophy, and law. Building on this interdisciplinary exchange, it formulates AI-specific audit areas for trustworthy use of artificial intelligence.

Fair behavior from the AI application towards everyone involved, adaptation to the needs of users, comprehensible, reliable, and safe functioning, as well as the protection of sensitive data are central criteria that need to be fulfilled in the trustworthy use of an AI application.

The audit areas presented here provide the basis for an AI audit catalog that we are concurrently working on. Neutral inspectors will be able to check AI applications for trustworthiness by using this catalog. With its many years of experience in the development of secure IT standards, Germany's Federal Office for Information Security (BSI) is an important partner in the drafting of this audit catalog. The certification system will allow us to make a major contribution to setting quality standards for artificial intelligence that is "Made in Europe", ensuring responsible use of technology, and promoting fair competition among different suppliers.

This white paper should contribute to social discourse on the use of artificial intelligence. After all, it is up to all of us to decide what the world of tomorrow will be like.

With this in mind, we hope you have an interesting and insightful reading.

Prof. Dr. Markus Gabriel
Professor for Philosophy at the University of Bonn

Prof. Dr. Dr. Frauke Rostalski
Professor for Law at the University of Cologne

Prof. Dr. Stefan Wrobel
Director of Fraunhofer IAIS and Professor of Computer Science at the University of Bonn

PRELIMINARY REMARKS

Executive summary

This publication forms a basis for the interdisciplinary development of a certification system for artificial intelligence. In view of the rapid development of artificial intelligence with disruptive and lasting consequences for the economy, society, and everyday life, it highlights the resulting challenges that can be tackled only through interdisciplinary dialog between IT, law, philosophy, and ethics. As a result of this interdisciplinary exchange, it also defines six AI-specific audit areas for trustworthy use of artificial intelligence. They comprise fairness, transparency, autonomy and control, data protection as well as security and reliability while addressing ethical and legal requirements. The latter are further substantiated with the aim of operationalizability.

Structure of white paper

The interdisciplinary approach to the topic is reflected in the chapter structure of this white paper. Chapter 1 provides an introduction to the topic and advocates the necessity of certification for artificial intelligence. In section 2.1, fundamental understanding of the functioning, possibilities, and limitations of the underlying technology is developed. The philosophical/ethical view of the problem, in particular the role of the ethical concepts of autonomy, freedom, and self-determination of people, is examined in section 2.2. The basics of the resulting legal requirements are discussed in section 2.3 with particular focus on responsibility, traceability, and liability for AI applications. Section 2.4 presents the effects of the different interdisciplinary perspectives, in particular with regard to the design of specific AI applications. In chapter 3, the specific fundamental audit areas are then justified and explained in separate sections from autonomy and control in section 3.1, through fairness, transparency, reliability, and security, to data protection in section 3.6. Finally, chapter 4 provides an outlook on the next steps planned in the development of a certification system.

Context

This white paper is the first fruit of an interdisciplinary project carried out by the KI.NRW competence platform that is aimed at developing a certification system for AI applications. This system will check for responsible usage from an ethical/legal perspective in addition to safeguarding the technical reliability.

1 INTRODUCTION

Each era brings its own challenges. We live in the age of digitalization. New technology is changing the way we live and work together on a massive scale. It is permeating almost every area of society – whether it be the world of work, road traffic, the healthcare sector, or simply the way we humans communicate with each other. Even if much of it is taking place silently or in an insidious way, the speed is unprecedented when compared with previous societal changes and would have caused fear and terror among our ancestors at the time of the Industrial Revolution in the 18th and 19th centuries.

One central driving force behind digitalization is the rapid development of artificial intelligence (AI) that was triggered by breakthroughs in so-called deep artificial neural networks on supercomputers. AI applications can even beat the top human experts in specialist areas like image recognition and complex strategy games. Artificial intelligence creates great opportunities for new technical applications, digital business models, and practical ways to make everyday life easier. Their applications are spreading unstoppably to a wide range of areas. Automated translation tools, voice control systems in homes, and self-driving cars are just some examples you will be familiar with. Artificial intelligence has a disruptive potential: the scientific and economic application possibilities are so far-reaching that it is currently hard to predict how our ways of perception and action will be changed by artificial intelligence. Furthermore, problem contexts will arise where we cannot sufficiently react with our traditional legal, political, ethical, and social means. AI research improves the generalizability of applications and their transferability to new contexts. Artificial intelligence is therefore gradually superseding older technology. Conventional value chains are being changed disruptively.

The increased productivity simultaneously relieves the burden on humans as they have to perform less monotonous or heavy work in certain areas.

It is generally expected that the number of AI applications will grow exponentially over the coming years. McKinsey predicts that AI could deliver additional global economic activity of around $13 trillion by 2030[1].
Furthermore, it is expected that artificial intelligence will contribute 1.2 percentage points to the annual growth of the global gross domestic product. The impact is therefore at least comparable with the productivity growth created by previous industrial breakthroughs, such as the steam engine (0.3 percentage points), industrial robots (0.4 percentage points), and the spread of information technology (0.6 percentage points). This impressive growth is down to more data being available and linkable, greater networking, and increasingly higher processing speeds. This all enables a greater degree of automation and individualization of products and services. In this area, individualization is more successful when more information is known about users[2] and customers.

It is obvious that, within a short time, the use of AI applications will have an impact on the way the whole of society lives and works together. This is particularly evident if we take surveillance systems as an example. For instance, facial recognition has also been tested in pilot projects in Germany such as at Berlin's Südkreuz railroad station. Among other things, however, the results were judged to be too erroneous. On the one hand, this shows that the question of reliability among AI applications sets new challenges compared with conventional software. On the other, however, it is now just a question of time and money before sufficient reliability can be achieved – at least in the aforementioned case of facial recognition in surveillance systems. In principle, AI-based intention recognition could also be combined with facial recognition so that it may even be possible to specifically set off an alarm when people act suspiciously in public places. The question immediately arises of how such surveillance – even if functioning optimally – would comply with current laws, and whether or how the law would have

1 Notes from the AI frontier: Modeling the impact of AI on the world economy, Discussion Paper, McKinsey Global Institute, September 2018, www.mckinsey.com/mgi

2 Gender-neutral pronouns have been used as far as possible.

to be changed for this. This leads to new ethical questions, since, on a societal level, we need to make fundamental decisions on which AI applications shall be permitted. Law and ethics need to cooperate in these new situations.

The scenario illustrates that forecast economic growth can only be achieved in the long run if sufficient trust is placed in AI technology. To create trust, AI applications need to be designed in a way that allows us to check whether they work safely and reliably. They also have to be in line with the ethical and legal framework. For this purpose, in addition to the technical safeguarding, we also need to clarify under which conditions usage is ethically acceptable, and what requirements result in particular from a legal aspect. The associated challenges involve basic issues that can only be resolved by an interdisciplinary team from IT, philosophy, and law. Since artificial intelligence is entering almost all spheres of society, the interests of numerous stakeholders are affected and deserve legal protection. A legal framework may need to be concretized or created for this purpose.

Conversely we do not, however, want regulations to be excessive and have a stifling effect on innovation. Likewise, they should not become outdated too soon and thus unusable due to the dynamics of technical progress. After all, ethics is not fixed for good which is why there is always the possibility of ethical progress and regression due to social and technological change.

Development of certification for AI applications

Since AI applications often work with particularly large quantities of data and use highly complex models, it is difficult in practice for users to check whether the assured features are being fulfilled. This is where a certification system for AI applications, which is based on professional and neutral checks, can create trust and acceptance – both at companies and among users and social stakeholders.

In view of the challenges posed by the use of artificial intelligence, the KI.NRW competence platform has set the goal of developing a certification system for AI applications, which can be employed by accredited inspectors.
In addition to assuring the technical reliability, the inspectors also check for responsible usage from an ethical/legal perspective. The certificate should confirm a quality standard that enables providers to design AI applications in a way that is lawful and ethically acceptable. It should also allow AI applications from different providers to be compared and thus promote open competition in artificial intelligence.

In addition to the requirement that an AI application must comply with ethical and legal principles, the interdisciplinary team identified six AI-specific audit areas that were defined in such a way that they could be evaluated as far as possible by individual specialists. The requirements of these audit areas are derived from existing ethical, philosophical, and legal principles (like for example, the general principle of equal treatment). They cover the areas of fairness, transparency, autonomy and control, data protection as well as security and reliability. While security covers the normal aspects of secure operation reliability concerns the special challenges set by checking complex AI models, like deep neural networks.

The question of how AI-applications can be used responsibly and reliably has been debated intensively by international social and scientific experts for a while now. On European level, the European Commission has set up a High-Level Expert Group (HLEG) for artificial intelligence. In April 2019, the HLEG drew up recommendations on which aspects should be taken into consideration during the development and use of artificial intelligence[3]. This white paper picks up on these recommendations, differentiates them, and goes a step further in some places. This is necessary because the recommendations from the HLEG are primarily of general nature. So far, they do not look at legal aspects (in particular, the specifics of individual national legal systems), nor at operationalizable ethical requirements with the clear aim of certification. In this respect, this publication takes a horizontal as well as vertical approach in comparison with the suggestions from the HLEG. In addition to the philosophical ethics, it examines law and puts the two in relation to one another. In order to meet the requirements of operationalizability, the audit areas determined in this way are also described more specifically and in greater depth than the HLEG categories.

3 https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

2 TECHNOLOGICAL, PHILOSOPHICAL, AND LEGAL PERSPECTIVES

2.1 Initial situation and questions from IT

In 1956, artificial intelligence was created as a branch of information technology with the aim of automating intelligent behavior. Inspired by cybernetics, cognitive science and neuroscience, a wide range of technologies were developed. They include intelligent agents that interact with the environment or each other via sensors and actuators, the combination of logic systems with heuristics, methods for symbolic knowledge representation and evaluation as well as machine learning (ML) with statistical processes and optimization, which have recorded massive growth particularly following recent developments. Soon after the discipline was created, people critically debated its responsible usage[4].

Many AI technologies are based on the use of models that contain knowledge about and experience of specific tasks. In machine learning, learning algorithms create the model from many examples, known as training data. There are calculation or "inference" processes for each kind of model that generate an output for an input. This allows the model to be then applied to new, potentially unknown data of the same type. Machine learning lends itself whenever processes become too complicated to be described analytically, but sufficient example data (for example, sensor data, images or text) are available. Using the learned models, we can make forecasts or generate recommendations and decisions – without any previously set rules or calculation methods.

Deep neural networks represent an important and large class of ML models. They consist of a large number of so-called artificial neurons created with software, which are linked to each other by means of weighted connections. This kind of network contains up to millions of open parameters that are optimized for the training data.

Structure of an AI application

The function of an AI application[5] is essentially determined by the trained-in ML models with calculation methods and possibly further pre- and post-processing procedures. This core in any AI application is called "AI component" henceforth. The AI component is always embedded in further software modules for the AI application. The modules activate the AI component and process their results further. They are ultimately responsible for the "behavior" of the AI application that is visible on the outside and for the interaction with the user. In particular, it is their responsibility to detect and work around AI component failure as well as react to faults and emergencies. An AI application can be stand-alone or form part of a system. For example, pedestrian recognition can be integrated as an AI application into an autonomous car, into a drone, or into a property surveillance system. In the discussion on and assessment of an AI application, one first important step consists of defining the limits of the AI application in the whole system and delimiting the AI components in the AI application.

4 Joseph Weizenbaum. Die Macht der Computer und die Ohnmacht der Vernunft. (German version of Computer Power and Human Reason) Suhrkamp Verlag, 1. Aufl. (1977).
Armin B. Cremers et al. (published on behalf of Association of German Engineers (VDI)). Künstliche Intelligenz. Leitvorstellungen und Verantwortbarkeit. VDI-Report 17, 189 S. 2. Aufl. (1993), VDI-Report 21, 121 S. (1994).

5 The following discussion places the focus on machine learning as a key technology for realizing artificial intelligence.

Challenges in the use of ML models

Dependency on training data

A specific request to an AI component returns the input for the ML model from whose calculation result the AI component generates an answer or reaction. Since AI applications "learn" their "behavior" from generalization of example data, the quality of the AI application greatly depends on the quality and the properties of the data stock used. If the training data are not statistically representative for the data that occur during operation, the results may end up being "biased" in one direction. For this reason, we need regular checks during operation to see how well the data distributions fit with each other and whether they diverge.

Probabilistic character

Due to the statistical nature of the model and qualitative uncertainties of input and learning data, the results are approximate and tied to more or less uncertainty. Often there is effectively not even a definite right or wrong answer. The AI component could output the best alternatives together with an uncertainty indication. When interpreting such results, a person needs to make a decision within their scope of discretion. In a fully automated application, corresponding provisions need to be made in the whole surrounding system to which humans essentially belong.

Comprehensibility and transparency of ML model and its results

Many ML models are so-called "black boxes". Black boxes are systems in which only the external behavior can be observed. The internal function mechanisms are not accessible, however, due to complexity or a lack of knowledge. It is therefore often impossible to understand how answers have come about.
For some applications, it may therefore be advisable to avoid certain types of ML models. We can, however, also supplement the ML model with another model, known as an "explanation model", that calculates which parts of the input were decisive for a certain result. For example, it has been discovered that an AI application in an image database recognized horses from a water mark, i.e. an artifact in the images, instead of from the shape of the animals.

Testing of ML models

Classic software test methods fail because the models can no longer be broken down into separately verifiable units. It is generally not even possible to find a formula to characterize reliable inputs. This was impressively demonstrated when an automatic traffic sign recognition system was totally confused by inconspicuous stickers affixed to the signs. Quantitative testing of the model using separate test data, which should have the same statistical distribution as training data, comes in here instead of modular testing.

Self-learning during operation

In principle, ML models can automatically carry on learning during ongoing operation, for example, by making use of user feedback. The ML model is then subject to constant change. One known example is the Tay chatbot from Microsoft that learned numerous racist phrases from its users within one day and was eventually shutdown. Since it is extremely difficult to set boundaries within which an AI component can continue to learn, the controlled use of such AI applications still represents an unsolved challenge at present. The currently best safeguard in this case is continuous monitoring of the AI application by humans.

2.2 Initial situation and questions from philosophy

Philosophy, in particular its sub-discipline of ethics, has now been assigned the task of providing ethics for artificial intelligence in order to counteract the disruptive potential of this technology. "Ethics of artificial intelligence" refers to a general requirement for how the application contexts (the field of application including the human/machine interaction), the technologies used, and the interfaces of the application contexts to the rest of the social and digital sphere have to be designed. The aim is for all participants to act well or be able to behave well according to their respective moral convictions and for nobody to be restricted in terms of rights, autonomy, or freedom. The certification of AI applications in their specific application contexts is an important first step towards general ethics for AI.

Two misunderstandings need to be avoided here: firstly, ethics of artificial intelligence refers, in this case, to specific AI applications for set tasks. This rules out questions like: which moral obligations and what responsibility do we have towards intelligent machines? Against this background, should we try at all to build artificial intelligence with general intelligence? When can an AI application count as a moral agent and does it possess freedom and rights? These questions do not concern the certification of specific AI applications which are actually involved at present.

Secondly, ethics for artificial intelligence cannot be implemented as a code in which every question that arises produces a binary yes/no answer from a specific problem context. The question of "which moral system can be programmed or modeled so that AI applications can be equipped in future?" is misguided. The reason is neither ethics can be conclusively programmed since they are in principle subject to change, nor can a consensus be reached on the correct moral system without running into difficulties. After all, ethics come from historically variable experiences gained by people. Societal transformations like digitalization give rise to previously unknown ethical problems. This means we first of all have to work out new guidelines by researching the specific human/machine interaction. These new guidelines have to be reconcilable with the universal value system of the human life form (human rights as a framework for law and ethics).

The main contribution of philosophy and ethics to the development of standards for artificial intelligence is thus newly defined guidelines for the use of our currently existing AI technology. These guidelines need to be in line with fundamental ethical key principles like human dignity, autonomy, and individual as well as democratic freedom. They set out the framework within which AI applications should move in their application context so that they do not contradict ethical basic principles like fairness or transparency. To this end, we have to look at both the AI application itself and its interface to the social sphere. This will only succeed if we place human AI users at the center.

2.3 Initial situation and questions from law

In terms of law, numerous challenges arise from the artificial intelligence technology that we have to face as a society. This includes the question of to what extent machine learning casts new light on the individual liability or the responsibility of persons, and thus the reason and criteria for "imputation". Systems that are controlled by machine learned models can have errors that could have a negative effect on individuals in particular in the form of prejudices. In addition, there is the difficulty that transparency is only possible to a limited extent with regard to learning systems. Whether, and if applicable in what scope, corresponding technologies are to be used in sensitive areas of society therefore requires clarification.

The healthcare sector is one example here. In this area, artificial intelligence provides support for doctors' work, for example, in cancer diagnosis or in the form of so-called "health apps". In healthcare, more and more robots will be used in future not least to replace human staff. As a result, artificial intelligence technology may change the healthcare market considerably in the next decade. The legal profession is also affected as is shown by advances in the field of "Legal Tech". Therefore, not least German courtrooms are being considered as another usage area for AI applications: an AI application could be used in court to make forecasts about the future risk posed by criminals. This system is already used in parts of the USA to help make court probation decisions. Furthermore, the use of digital technology in the fight against crime seems to be less futuristic as it has already become reality for the German police. The term "predictive policing" refers to the usage of data to predict criminal activity. This system is used for police operations planning. "Predictive policing" is one application area that is earmarked for expansion in the "Artificial Intelligence Strategy of the German Federal Government".

The question of how we want to live in our society is at the core of all of these developments. Is there a "human image of digitalization" – and can this be reconciled with

a free state under the rule of law? People endowed with human dignity are the focus here, for which Art. 1 Para. 1 of Germany's Grundgesetz, the Basic Law, provides the normative basis. According to the Grundgesetz, people may not be degraded to mere objects of state actions (see for example German Federal Constitutional Court decision 9, 89; 27, 1; 28, 386; 117, 71, 89; 131, 268, 286 subsequent to G. Dürig, Public Law Archive (AöR) 117 (1956), 127). The protection of this right requires particularly critical examination in times of disruption by artificial intelligence, which premises intensive cooperation between law and philosophy. One thing is definite here: technological revolutions should not be understood as "no-brainers". Moreover their course of events lies in the hands of people since they are the essential players. From a legal viewpoint, the KI.NRW certification project therefore pays attention to the configuration possibilities that are available when it comes to use of artificial intelligence. In this way, we aim to make a relevant contribution to the image that society will create of itself in times of major technological advances in the field of artificial intelligence.

2.4 Interdisciplinary observations

To adequately account for disruptive technology, which, like artificial intelligence, works itself into the roots of a society and can cause changes on a previously unknown scale and at unforeseen speed, we need to use a holistic approach. In a free state under the rule of law, people are at the center of philosophy, law, and technology. The cooperation of sciences is therefore not only abstractly desirable, but is also an important requirement of our times.

Design of the ethical/legal framework of artificial intelligence

Our society and thus each individual has the possibility to (help) decide how the world, in which we want to live with artificial intelligence in the future, should look. Philosophy, law, and technology play a central role in the discourse that has to be conducted for this purpose. Technological development generates the problem areas of this social discourse. At the same time, it shows what is actually possible and what belongs to the realm of science fiction. Philosophy reassigns central terms of ethics, such as the moral figure, in the context of artificial intelligence and provides reasoning for the universal validity of certain ethical principles and legal norms like, for example, human rights. The framework for a useful and purposeful societal discourse is set only through this process. Law uses ethical arguments in implementing the outcome of the discourse to find the legally correct solution. This is above all relevant in areas where law still has not arrived on the scene. In view of the large number of changes that artificial intelligence applications cause in each area of society, the question arises from the lawyers' viewpoint of whether there is a need for regulation. Furthermore, basic legal terms are put to the test when confronted with new technical developments. This concerns, for example, the term for responsibility or "guilt". Basic terms from philosophical ethics like justness, equality, autonomy, fairness, and transparency etc. are also affected by this. These need to be conceived precisely for the context of AI applications as these terms gain a specific meaning that they only get through the new technology. These meanings can only be clarified in trilateral collaboration. The question arises in the context of artificial intelligence as to whether we can keep the previous terms or whether they need modifying. In this case, any legal evaluation requires clear understanding of the technical contexts. Above all, this concerns the actual possibilities of AI applications and thus the question of effective implementability of legal requirements. If these do not exist, a situation occurs where something is demanded from the legal side that cannot be achieved technically (for example, unlimited transparency). If, however, a corresponding technical feasibility does not exist, the further legal question arises of whether both permissibility of the respective AI application can be justified.

Development of specific AI applications

In the design of AI applications within an existing ethical and legal framework, it is essential to incorporate the viewpoints of all three disciplines. As early as the design phase of the AI application, it must be clarified whether the

application is ethically and legally permissible and, if so, what boundaries should be set for its configuration. One necessary criterion in this case is to give all those involved the same possibilities to make a moral decision, which they would also have if AI was not used, and to observe their rights and freedom. Many other subsequent questions that result from this – for example, what fairness means in the context of the application or what effects on the user, such as emotional ties to the AI, are acceptable – cannot be answered from a technological perspective alone, but instead require a holistic approach again.

If the general permissibility of the AI application has been ensured, interdisciplinary questions also result for the further development up to release, e.g. as Open Source. These questions concern handling inevitable conflicts and trade-offs between the different audit areas. A different balance between the individual values is required in respective different contexts. Conflicting interests can be brought into a balanced relationship with each other through the ethical/legal principle of proportionality. In this way, all perspectives of the players involved are incorporated into the necessary weighing of interests. Although weighing decisions cannot be made on the meta level in individual cases, the proportionality principle provides an instrument to establish the reliability of specific AI applications.

3 AUDIT AREAS FOR A CERTIFICATION SYSTEM

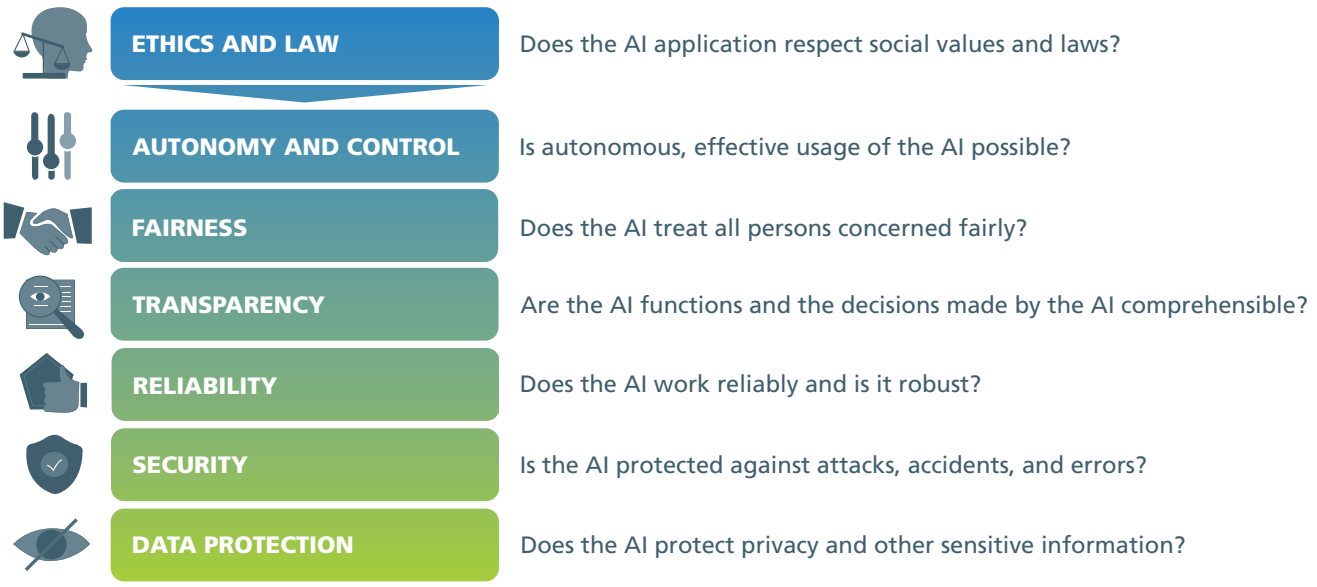
Due to their disruptive potential, it is particularly important for AI applications to guarantee concurrence with the philosophical, ethical, and legal framework. Their certification serves above all the protection of the fundamental legal and ethical interests of people. We should be able to avoid impermissible adverse effects on individuals or groups in this way. In this respect, AI certification follows the general purpose of averting injustice or ethically unjustified conditions in society. In addition to the rights to freedom of individuals and the principle of equal treatment, this concerns in particular also general social interests like, for instance, the protection and preservation of the environment as well as the constitutional democracy.

A large number of substantiations can be derived from these basic values and principles of a liberally organized community taking the constitutional principle of proportionality into consideration. In this way, audit areas that are significant for certification are drawn up on the basis of ethics and law as well as IT requirements.

For the development of an AI application, this entails that the application area, purpose, and scope as well as affected persons need to be identified at an early stage. All players who are directly or indirectly affected should be involved appropriately in this process. A risk analysis should be performed that covers the possibilities of misuse and dual use whose consequences need to be included appropriately in the further development. Finally, the application should "by design" be built in a way that it can be audited and tested to the defined extent.

3.1 Autonomy and control

Autonomy is recognized both as an ethical and legal value. In terms of philosophy, autonomy forms the basis of all values since, as a human community, we have to give ourselves values. It is generally the capability for morally relevant self-determination. This justifies the freedom of individuals to make autonomous decisions. This also covers all decisions that relate to your own legal position and moreover the freedom to determine the goals of your own actions as well as to choose the means to reach these goals.



AI applications are taking over more and more routine activities and are increasingly autonomous. We need to note that mobile systems (for example, robots and vehicles), which are controlled by the assigned AI applications, are also often described as "autonomous". However, these systems are only capable of choosing the means and not the actual goal. In this context, we misleadingly say that the system possesses "autonomy to act" where in fact this results from the human goals. For this reason, an area of conflict with autonomy (of people) also results as such AI applications can influence people in the choice of their goals and means. The latter is particularly the case if the AI application interacts with human decision-making by, for example, generating decision suggestions, creating control commands (and possibly executing them), communicating with the people directly (virtual assistants, chatbots,...), or being integrated into work processes.

Artificial intelligence may not disproportionally restrict the autonomy of individuals and social groups. Against this background, it is important in the development and operation of an AI application to state to what extent individual or collective users can develop excessive trust in the AI application, build up emotional ties, or be impermissibly impaired or directed in their decision making. The distribution of tasks and interaction possibilities between the AI application and user therefore need to be clearly and transparently regulated. Users need to be familiarized with the possible risks related to potential impairment of their autonomy, and with their rights, obligations, and options to intervene as well as make complaints. The user must be given the possibility to control the system within an adequate scope. Also, they must have the option of revoking their consent to using an AI application. This should not simply be a yes/no option for the user, but instead multiple usage possibilities should be provided. In particular, it must also be possible to switch off the application completely. Users need to be supported adequately in the safeguarding of their autonomy by receiving the necessary information about the behavior of the AI application during operation without being overwhelmed. The latter in particular also needs to cater for people with special needs. In addition, appropriately secure intervention possibilities need to be provided in case a risk to the autonomy of the user is detected.

3.2 Fairness

Emanating from the general principle of equal treatment, safeguarding the principle of fairness is to be required from an AI application both in an ethical and in a legal respect. This refers to the ban on treating the same social issues unequally or differing ones equally unless a different procedure would be objectively justified. The principle thus stretches to the ban on unjustified discriminatory treatment in an AI application and rules out impermissible discrimination. This means in particular that individuals may not be discriminated against again in the social result due to their affiliation to a marginalized or discriminated group. For example, people with certain surnames, a specific religious affiliation, or a specific gender may not be given a better or worse evaluation. Also, voice control systems must be able to react to people with specific accents or sociolects and be customizable. Furthermore, facial recognition software may generally not make more frequent errors with people of a certain skin color or other phenotypical features.

AI applications learn from historical data. These data are not necessarily free of prejudice. If, for example, the data discriminate against women, the AI component can thus adopt these prejudices, too. Also, certain groups may be underrepresented in the data basis. We then talk of bias. Bias can also lead to decisions that are unfair. One known frightening example is the incorrect classification of people with dark skin as gorillas by Google Photos. Representative training data must therefore be provided. Furthermore, an improvement in the output of the ML model comes into consideration as a suitable instrument for avoiding bias.

To operationalize fairness, a quantifiable fairness term needs to be respectively developed from a technical viewpoint. The groups that should not be discriminated against therefore need to be identified in a first step. These groups can be social minorities, but also companies or general legal persons as is the case, for example, with pricing on digital marketplaces. In a second step, the chosen fairness definition needs to be quantified. The differentiation of group fairness and individual fairness should be highlighted in particular in this process. Regarding group fairness, it should be required that the results for all existing groups are comparable, for example, in the sense of "hit probability" in all groups. In the

case of individual fairness, the same treatment of the same individuals is set
as a standard.

3.3 Transparency

The transparency of an AI application can be decisive for its acceptance. Two aspects should be differentiated in this area. Firstly, information on the correct usage of the AI application needs to be available. Secondly, it is about requirements for the interpretability, traceability, and reproducibility of results that require insight into the internal processes of the AI application.

Information on usage of an AI application

First of all, it has to be generally clear in a communication situation that the communication is occurring with an AI application. Furthermore, the players need to be adequately familiarized with the use of the application. This includes understanding what purpose the application has, what it does, what the potential risks are (also in terms of other audit areas, for example, reliability, security, and fairness), and who the target group of the application is.

Traceability and interpretability of ML model

From an ethical/legal viewpoint, a conflict of interests can occur between the desire for transparency for the user (or for interested groups) on the one side and the safeguarding of trade secrets or the general social safety on the other. This results specifically in the following requirements for the transparency of an AI application:

- AI applications that affect the rights and interests of third parties generally must be transparent. Transparency means the traceability of how the AI application works.

- AI applications do not need to be made transparent to the outside. This does not apply if there are predominantly social interests for the understandability of the AI application.

- AI applications that affect the rights and interests of third parties may be non-transparent in exceptional cases if this is proportionate when the conflicting interests are weighed up.

This second type of transparency concerns the internal processes of the AI application and in particular of the ML model. This involves the questions of interpretability, traceability, and reproducibility of results for different players and purposes. In particular, the following should be demanded among other things:

- Users must be able to comprehend the output of the AI application to the extent that they provide informed consent or refusal. This can frequently occur by showing the passages relevant to the decision during input.

- The information provided needs to be selected in a way that users are not overwhelmed with irrelevant details to allow informed intervention when using an AI application.

- Experts generally need to be able to trace the functioning of the AI application on technical detail level, for example, for the purpose of improvement or clarifying conflicts. The experts do not have to be able to predict each output of an AI application. However, its general behavior in principle needs to be explainable, traceable, and documented during development and also later in productive operation. Logging, documentation, or archiving of the design, data, training, testing/validating the model as well as the embedded environment are used for this purpose.

From a technical viewpoint, the question of general transparency is not trivial and the field of tension between greater accuracy or robustness and the explicability of models is a long-known dilemma in the world of AI. In many cases, "black box" models are actually more accurate and more robust than, for example, rule-based models, but they can only be interpreted to a limited extent. This explicability can partly also be achieved through subsequent processes, for example, by training explanation models or analyzing the input/output behavior of models (known as LIME analysis – Local Interpretable Model-agnostic Explanations). The interpretability of models is currently an active research field and great efforts are being made to understand the learning processes of "black box" models better as well as to visualize their internal processes and explain the resulting decisions.

3.4 Reliability

From a technical viewpoint, reliability represents a collective term that partly compromises clearly different aspects of the quality of an AI component: the correctness of the AI outputs, the estimate of the ML model uncertainties, and the robustness to harmful inputs (e.g. adversarial attacks), errors, or unexpected situations. New kinds of error modes, which are not typical for humans and thus unexpected, can lead to situations that are potentially critical because they have not been practiced – in particular in direct human/machine interaction.

In-depth knowledge of the application is required to evaluate these reliability dimensions for a specific AI application and to define under what requirements the application can be classified as reliable according to these dimensions. For this classification, the requirements that have already been collected, the initial risk assessment as well as the ethical and legal framework should be fully taken into consideration. The conversion of the requirements into quantitative measurements and target values requires knowledge of the domain as well as mathematical/technical expertise and is never complete by nature. The same applies to the description of the application area of the AI application. It should be specified as precisely as possible and be formalized in order to ensure that the training and test data used sufficiently cover the inputs to be expected during use of the AI application. In any case, the reliability of the AI application should be configured for the capabilities of the people using it.

Correct implementation of the training routines and of the finished trained model is an essential factor in meeting these requirements. The tests to be performed for this should be established in the area of machine learning and be configured for the respective application. If ML model weaknesses are uncovered, we need to react to this with suitable correction mechanisms right up to use of a back-up plan. The reliability of the AI application should be guaranteed at all times in productive operation in this case.

This implies that the correct function needs to be checked at appropriate regular intervals. In order to also increase the reliability gradually, suitable measures should be established, for example, by saving challenging scenarios in productive use.

3.5 Security

Security in the sense of protection against attacks and safety in the sense of protection against dangers arising from the AI application are at least just as highly important as with other information and technical systems. The security and safety concepts can be used on the whole AI application in which the AI component is embedded and on the AI component itself. AI-specific risks should be intercepted or handled in a suitable way. These risks can come in the form of function failure or major function changes in the AI component as well as unauthorized information leaks. The causes of the function failure, for example, adversarial attacks, and major function changes are already reacted to inside the AI component, which comes under the reliability audit area. If this is not possible to the full extent, the measures of the surrounding AI application are effective and the responsibility lies with the security audit area.

The HLEG has defined abstract security goals for AI applications. These abstract goals (and those going further) are, however, far off being put into operation for instance through an audit catalog or a standard. Conversely, a whole series of operatively verifiable specifications and standards do exist particularly in the area of security. However, they do not refer specifically to the special features of AI applications. The aim of the security audit area is to gather the requirements from existing standards that are indispensable for protection against attacks and the dangers of AI applications and supplement them with further specific AI requirements.

3.6 Data protection

AI applications are suitable for intervening in a large number of legal positions. Quite frequently, this involves interference in the private sphere or the right to informational self-determination. For example, AI applications often process sensitive information, for example, trade secrets, private information, or personal data, such as voice recordings, photos and videos. Therefore, we need to ensure that the relevant data protection-law regulations, for instance, the General Data Protection Regulation (GDPR) and Germany's Federal Data Protection Act are observed. AI applications may not just pose a risk to the private sphere of individuals. Furthermore, they could affect (trade) secrets that do not contain personal data as defined by the GDPR, but require protecting in terms of ethics and law. This can involve, for example, machine data that contains information on the process utilization

or error rates regardless of which person was active as the machine operator. The AI certification should ensure that data protection risks and measures in the AI application are analyzed and documented sufficiently by the AI system so that the data protection officer, who generally has to be nominated, receives useful support. This will enable them to conduct the investigation and make the final decision or approve the data protection while taking the special data protection challenges related to artificial intelligence into consideration.

The challenges for data protection are potentially greater than in traditional IT systems since AI applications often gather together data that were previously not linked. Also, new methods of linking data are not formed until machine learning is employed. The more data are linked, the greater the risk that people or, for example, specific business premises can be identified even without direct specification of corresponding attributes. For example, it is possible, with approximately 95 percent reliability, to re-identify persons from the way they use a computer keyboard. If there were a public (or commercially available) database that assigned keyboard typing patterns to persons, the typing sample would become a so-called "quasi-identifier" that enables a link to a person to be established. Also, AI methods can potentially create references to persons during the processing of text, speech, and image data as well as logged usage data. In addition, there is a risk that a trained model re-allows references to a person without actually containing personal data.

This results in the gathered information having to be effectively protected both during training and also during operation. AI applications may access personal data only with the consent of the owner. Further processing and disclosure to third parties may – subject to further restrictions – occur exclusively with consent from the owner of the legally protected right. It must be ensured that there are no gaps in the protection that enable unauthorized access. Individuals must be granted the option of deleting their data. The necessary measures thus include notifying the person concerned about the purpose and use of personal data or data derived from them and providing adequate consent, inquiry, objection, and revocation mechanisms related to the use of personal data. Compliance with the principles of data economy and use for a specific purpose should be indicated as well. A risk analysis should also be performed to examine the potential to produce a reference to a person. This analysis should check any measures taken to make data anonymous or to aggregate data against the potential for re-identification via links to background knowledge.

4 OUTLOOK

This white paper is the first fruit of an interdisciplinary project carried out by the KI.NRW competence platform that is aimed at developing a certification system for AI applications. This system will check for responsible usage from an ethical/legal perspective in addition to safeguarding the technical reliability. An AI audit catalog, which is currently being developed and will allow accredited inspectors to assess AI applications in a professional and neutral manner, will form the basis for certification.

The plan is to publish an initial version of the audit catalog at the beginning of 2020 and then begin with the certification of AI applications. Due to the complex nature of the topic, the first version will set suitable restrictions related to the applicability in several places, for example, in the area of further learning during operation or for the control of critical security applications. A series of different AI applications is used during the development of the audit catalog to check the integrity and universality of the audit objectives and requirements. We will also evaluate and demonstrate the use of the catalog as part of this.

One special task in this area will be checking the audit objectives against existing standards and the differentiation from existing audit catalogs and laws, for example, for IT security and the General Data Protection Regulation. For this reason, the project team is collaborating with Germany's Federal Office for Information Security (BSI) to incorporate their many years of experience in the area of IT security for the drafting and recognition of IT testing standards.

The methods and possible applications of artificial intelligence are being continuously developed on a massive scale. We can assume that society's concept of ethics and the regulation of artificial intelligence will be shaped with them. Therefore, the audit catalog must be a living document that undergoes continuous updating from the areas of information technology, law, and philosophy. At the same time, the scope of the catalog will be expanded step by step, and special catalogs will be produced for certain application areas and risk classes.

5 IMPRINT

Publisher

Fraunhofer Institute for Intelligent Analysis
and Information Systems IAIS
Schloss Birlinghoven
53757 Sankt Augustin
Germany
kinrw-pr@iais.fraunhofer.de
www.iais.fraunhofer.de

Contact

Dr. Maximilian Poretschkin
Telephone: +49 2241 14-1984

Cover photo

© mila103, ryzhi, zapp2photo / fotolia.com

Layout and design

Svenja Niehues, Fraunhofer IAIS, Sankt Augustin