Lernende Systeme

GERMANY'S PLATFORM FOR ARTIFICIAL INTELLIGENCE

# Criteria for Human-Machine Interaction when using AI

## Approaches to its humane design in the realm of work

SPONSORED BY THE

Federal Ministry of Education and Research

acatech
NATIONAL ACADEMY OF SCIENCE AND ENGINEERING

WHITE PAPER

Norbert Huchler et al.
Working Group Future of Work and Human-Machine Interaction

# Table of Contents

# Executive Summary

Artificial Intelligence (AI) offers a wide range of potentials for safe, autonomous and self-determined work as well as attractive and competitive jobs. For example, AI-based assistance systems can relieve workers of strenuous or dangerous tasks and support them in complex processes and decisions. At the same time, AI systems are changing the interaction between humans and technology in the realm of work. In the future, humans and machines will interact even more strongly – and in different ways – than in the past since Machine Learning (ML) and similar technologies enable machines to perform certain tasks independently and to learn continuously in the process.

The increasing collaboration between humans and technology makes it necessary to readjust the distribution of work when using Artificial Intelligence. In order to shape this collaboration in the interest of humans, technology must use the advantages and potentials of human thinking and acting as a starting point and must focus the interaction on mutual support – rather than on replacement or conflict. A coordinated balance that does justice both to workers and to the technological and economic potentials of Artificial Intelligence increases the chances for individually and socially accepted use of Artificial Intelligence in the realm of work.

Clearly defined guidelines for the new distribution of tasks are necessary to create safe jobs, train qualified workers, and implement design for good and humane work. That is the starting point for this white paper. It presents a set of criteria for human-machine interaction in the context of work. The objective of the criteria is design of the human-machine interaction that is future-oriented and human-centered long-term.
The criteria can be grouped in four clusters:

**Cluster 1: Protection of the individual**

- Protection of safety and health
- Data privacy and responsible performance monitoring
- Diversity sensitivity and non-discrimination

**Cluster 2: Trustworthiness**

- Quality of the available data
- Transparency, explainability, and consistency
- Responsibility, liability, and trust in the system

**Cluster 3: Reasonable division of work**

- Appropriateness, relief from strain, and support
- Agency and situation control
- Adaptivity, error tolerance, and customizability

**Cluster 4: Supportive working conditions**

- Scopes for action and richness of work
- Conducive to learning and gaining experience
- Communication, cooperation, and social embeddedness

These criteria are addressed to actors involved in planning and developing self-learning systems as well as actors involved in implementing AI systems in companies. The set of criteria is intended to provide guidance for designing the division of work between humans and technology when applying self-learning systems. In addition, the criteria are intended to inspire actors to develop existing regulations further – for example in standardization, legislation, or industrial relations – and to enable more flexible, self-determined, and autonomous work in the future.

# 1. Division of work between humans and machines

The use of Artificial Intelligence (AI) at work requires a readjustment of the "division of work" between humans and technology. The reasons for this include the redistribution of activities in work systems in general, but particularly the increasingly close cooperation and direct collaboration at the human-machine interface due to learning systems: machine learning and similar technologies strengthen the ability of complex technical-organizational systems to assume the role of an "actor"[1] in interacting with humans and to gain agency for certain tasks in that interaction.[2]

This new division of work and roles requires specific criteria that formulate starting points for designing human-machine interactions in the realm of work that are future-oriented and human-centered long-term. In this context, systematic analysis of the different potentials of human and technology on the one hand and the different interests of the actors involved on the other are important.

This kind of "division of work" is all the more successful the better the respective strengths and potentials of human and technology – i.e. the specific human and technical capabilities and characteristics – can be related to each other in a mutually reinforcing, "co-evolutionary" way (Huchler 2016). Sensitivity to the opportunities and necessities of complementary collaboration between human work and semi-autonomous intelligent or self- learning systems is needed so that Artificial Intelligence does not weaken humans in their central role in the realm of work, but rather validates and strengthens them.

A realistic examination of the specific advantages, but also the immanent limits and deficits inherent in the technology and its application – in the case of Artificial Intelligence, for example, data dependency, its lack of "feel" for causalities or lack of their representation, the danger of circular reasoning or path dependencies – is important when developing design criteria.

If human-machine interaction when using Artificial Intelligence is to be humane, it is important that the technology is compatible with human action, especially precisely where the particular advantages and potentials of human thinking and acting lie – such as acting under uncertainty or with incomplete information and contradictions, the combination of

1 Since increasingly extensive interaction skills are "inscribed" or "programmed" into learning AI systems and more complex divisions of work are possible at the interface, the "social actor" is faced with a "technical actor." However, a strict distinction must be made between these types of actors – for example regarding their competencies, legal implications, and ability to bear responsibility.

2 In the following, "interaction" includes not only "social interaction" between individuals, but also "social interaction with objects" as well as "interactions inscribed in objects" and "object-mediated interactions." In addition, "agency" pertains to formally defined tasks/functions, and individuals, organizations/institutions (collective actors) as well as technical objects can possess agency.

specialist knowledge, experience, and implicit knowledge to form competencies and experiential knowledge, or even the capability to attribute meaning to information depending on the social context in question ("indexicality"). This places high demands on the interactive compatibility or "complementary adaptivity" (Huchler 2019) of the technical system – in order to place mutual support at the center of interface design rather than replacement or conflict.

The potentials and requirements of human-centered development of technology and work design must be negotiated and coordinated with the demands and opportunities of economic rationalization and technological progress. The applicable legal provisions – such as labor law, occupational safety laws, protection of personality rights, regulation of data privacy, and worker participation – constitute the background and starting point. A coordinated way of balancing that does justice to the workers and their development on the one hand and to the potential of technological innovations for humans and society on the other will enable more appropriate and effective use of Artificial Intelligence in the workplace (including the social consequences) as well as in Germany in general.

A design perspective that achieves an optimal "division of work" at the interface between humans and technical systems on the basis of the different competencies will offer a realistic view of application scenarios, promote the practical applicability and sustainability of developments, and, not least, could help to address reservations and concerns and promote social and technical innovations.

Developing criteria for designing human-machine interaction in self-learning systems is an important element with respect to introducing these systems into the realm of work. Suitable design concepts are an essential element for implementing transformation processes in companies. Application scenarios in which the "division of work" is appropriate to the design of human-machine interaction must already be reflected upon when developing technical systems – and later when using self-learning systems.

A participatory course of action that considers industrial relations is necessary with a view to acceptance on the part of workers as well as a good fit to the work process in question. This offers a framework for the design and introduction of AI technologies and self-learning systems that is jointly supported and promotes innovation.

This white paper, which was prepared by the Working Group Future of Work and Human-Machine Interaction of Plattform Lernende Systeme, focuses on developing criteria for the design of the human-machine interaction in the context of work. The paper particularly addresses actors involved in planning and developing self-learning systems as well as actors involved in implementing AI systems. The objective is to ensure that the following set of criteria is considered early on in technology-driven innovation processes relating to Artificial Intelligence, for instance as a tool for reflection.[3]

3 The future design instrument follows other models that can be used as starting points for developing criteria for the design of human-machine interaction – for example, the MEESTAR model (**M**odell zur **e**thischen **E**valuation **s**ozio-**t**echnischer **Ar**rangements, model for ethical evaluation of socio-technical arrangements) (see Manzeschke et al. 2013).

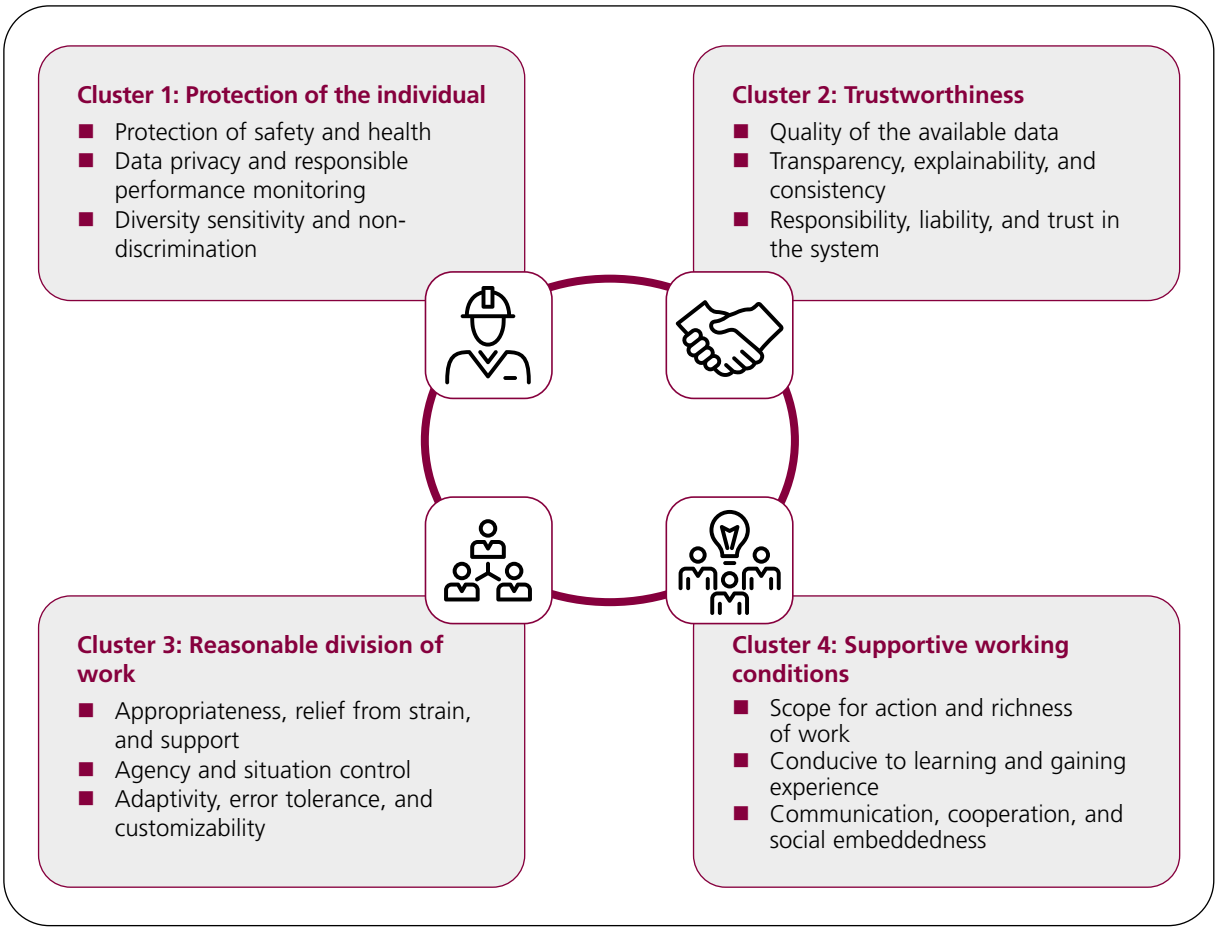
# 2. Designing human-machine interaction

The objective of the set of criteria is to develop approaches specifically for designing human-machine interaction when using Artificial Intelligence and to go beyond existing approaches for designing digital technologies. One important distinguishing feature is the changed "division of work" and interaction at the interface between humans and AI-based assistance systems, which involves one or more transfers of agency between humans and machines or even an overlap of their agency.

The set of criteria is intended to provide support for designing, both in technical and workplace-related terms, the "division of work" between humans and technology when using Artificial Intelligence, and to inspire further development of existing regulations – for example in standardization, legislation, or industrial relations. The starting point and framework are to be found in existing law and regulations on industrial relations – such as provisions in the law concerning labor, occupational health and safety, personality rights and data privacy, or legal regulations on worker participation. At the same time, the criteria are also intended to support the goals of a site's productivity and innovative capacity and enable more flexible, richer and more self-determined work.

In the following, the set of criteria is first presented as an overview (Figure 1). The individual criteria are subsequently presented in detail. The criteria can be grouped into four clusters: the requirements for the protection of the individual, for example concerning safety and privacy (Cluster 1) and trustworthiness (Cluster 2), requirements for workplace design with regard to reasonable "division of work" (Cluster 3), and supportive working conditions (Cluster 4). It should be noted that there are numerous interlinkages between the criteria, which makes it difficult to delineate them clearly in all cases. In addition, it is important to consider the different levels of automation and autonomy as well as the levels of criticality of the application in question when developing and using AI systems in companies. Each of the criteria must be set in relation to the various levels and mutually coordinated with them (similarly to other areas such as mobility or industry: acatech 2016, acatech/Fachforum Autonome Systeme 2017, Plattform Industrie 4.0 2019b). The focus is on interactive human-machine interfaces with humans and the technical system reinforcing each other.

**Figure 1: Overview of clusters and criteria for human-machine interaction in the context of work**




## 2.1 Cluster 1: Protection of the individual

The safety of self-learning systems and the protection of the individual against risks and negative consequences of the use of Artificial Intelligence are crucial elements for its use and acceptance in the working environment. They should already be considered during the development phase. In addition to the various requirements of occupational health and safety, this is also a matter of complying with the laws protecting workers' personality rights, ruling out unjustified performance monitoring, and protecting against discrimination of workers. The following three design criteria provide key starting points for protecting the individual when using and applying Artificial Intelligence.

## Criterion 1: Protection of safety and health

**Objectives:** Promoting the health of workers; avoiding risks to physical and mental health; protecting against accidents and damage (personal injury and damage to property) and avoiding negative consequences of physical or psychological stress; providing protection by minimizing risks and wrong decisions.

**Starting points:** AI-based assistance systems are used in different areas – therefore the priority objectives in the context of AI regarding protection of safety and health vary as well.

In the case of mechanical industrial systems, the focus is on preventing accidents and damage (personal injury and damage to property). In this context, it should be noted that suitable design of the human-machine interaction can prevent negative consequences of physical or psychological stress (e.g. monotony or mental saturation) and thus contribute to preventing accidents and damage.

Concerning knowledge and service work, which relates to information and interaction, the main focus – from a safety perspective – is on protection against financial and professional risks, damage to reputation due to planning errors inherent to the system, and wrong decisions and actions with negative consequences. In terms of protection of health, this is primarily a matter of avoiding psychological distress at the human-machine interface. It should be noted that these systems do not produce a final service or effect, but that they are used in particular to support decisions and actions.

The overall purpose is to link aspects relating to safety and security with humane design of AI-based work systems, thereby preventing risks to the mental health and well-being of workers (e.g. through excessively overburdening or unexpected behavior of AI assistance systems).

## Criterion 2: Data privacy and responsible performance monitoring

**Objectives:** Protecting and strengthening the personality rights of workers; data economy and purpose limitation of data use in Artificial Intelligence; legally compliant, responsible, company- and industry-specific use of the opportunities for performance monitoring agreed with worker participation; avoiding data analyses to monitor individual performance or behavior; development of a positive culture of performance feedback supplemented by "analog" means; transparency concerning data analyses and their use, and empowering workers to handle data transparency.

**Starting points:** Collecting data of sufficient quality and analyzing and evaluating them with the help of improved technical solutions and methods are at the core of self-learning systems. This increasingly also involves sensitive behavioral and performance data. At the same time, it is becoming ever easier to collect, store, and evaluate this data.

In the work context, data analyses can in principle be used as safety technologies, to improve product quality and processes, to prevent errors, for assistance, for relief from strain, and partly for prevention – but also for performance control and monitoring. A distinction is to be made between aggregated and individualized data. For example, personal profiles and usage data can facilitate work processes for individual users and, at an aggregated level, contribute to enhancing quality and efficiency for the company. If there is no such separation, conclusions can be drawn from analyzing personal user profiles which in extreme cases can lead to "transparent" workers and continuous individual performance monitoring. Monitoring workers' behavior and performance should be avoided. This applies particularly to predictions based on workers' personal data. In addition, it is precisely the methods of predictive analytics based on Artificial Intelligence that cause the typical problems of false conclusions, discrimination, and indirect influence on behavior.

Since both data access and the use of analytical methods are becoming easier, a responsible approach to the topic of performance monitoring is becoming increasingly important. Specifically, this implies protecting and strengthening personality rights as a basis, especially in the form of defining the purpose of AI systems and considering it in a differentiated manner as well as the data processing required for this purpose (e.g. improvement of processes, further development of workers, protection of workers from excessive burdens), and avoiding undifferentiated monitoring and use of all available performance data. With respect to data quality, it also makes sense to collect only the data required for an application (data economy). Industry-specific and company-specific differences must be taken into account in this context.

First and foremost, the development and use of Artificial Intelligence must comply with the laws protecting workers' personality rights, from the General Data Protection Regulation (GDPR) to labor law and laws on worker participation. For example, it is impermissible as a matter of principle to process biometric data in order to uniquely identify workers. In terms of designing the human-machine interaction, this means implementing trustworthy procedures as well as safeguarding regulations and also taking data privacy regulations into account even during the technical design phase of the systems, if possible, or strengthening data privacy through their design. For example, cameras can make faces anonymous locally even before transmitting the image data. Data can also be linked to delete functions (e.g. after usage or a certain amount of time), and access rights and usage possibilities can be restricted.

For which purpose and how the collected and analyzed data should be used is becoming increasingly important even during the design phase of systems and then also during use of the systems in the company. Accompanying measures (e.g. leadership, education, building trust) as well as supplementary regulations are necessary beyond the technical design of human-machine interaction in order to protect personality rights and data privacy.

## Criterion 3: Diversity sensitivity and non-discrimination

**Objectives:** Protecting against discrimination of individuals or groups and avoiding possible distortions; existing legal system as a basis for Artificial Intelligence displaying diversity sensitivity and non-discrimination.

**Starting points:** Discrimination as unjustified equal or unequal treatment must be differentiated from factually explained distinctions that are justified or even necessary for desired and accepted applications and functionalities.

Discrimination can arise or be reproduced and intensified for different reasons when using Artificial Intelligence and self-learning systems: for one thing, societally established prejudices can be perpetuated (preexisting bias) by (self-) learning systems – e.g. by algorithms or training data; for another, technical falsifications – e.g. in sensor technology – can lead to discrimination (technical bias); furthermore, the interaction between software and application may result in unjustified equal or unequal treatment (emergent bias).[5]

The existing legal system is the starting point for assessing and evaluating whether discrimination occurs in human-machine interaction. It is also the basis for an appropriate and necessary analysis of the circumstances and reasons. Avoiding preexisting technical and/or emergent bias requires sensitivity to diversity in the development and application of AI systems as well as careful selection of training methods and data. Furthermore, people should be aware that Artificial Intelligence and self-learning systems do not necessarily make more neutral or objective decisions than humans do (Plattform Lernende Systeme, 2019).

4 Although it is true that "justified discrimination" or differentiation is often necessary in order to obtain correct results from AI systems – such as in medical diagnoses by differentiating between women and men (see Plattform Lernende Systeme, 2019), the present paper follows the established definition of discrimination, namely always factually unfounded equal or unequal treatment.

5 Conversely, self-learning systems are dependent on (well-considered) bias in order to be able to learn at all (i.e. "learning bias," which makes it possible to infer from given data to other data). The issue of discrimination is about avoiding ethically unjustified or socially undesirable bias.

## 2.2 Cluster 2: Trustworthiness

If workers are to trust AI systems, manufacturers and companies using them must prove themselves trustworthy by designing the technology to be human-centric in concrete terms. This means that when people experience the human-machine interaction, their trust can be built up step by step or diminished abruptly. This applies both to trust in the technical system and trust in the companies involved in its manufacture and use, as well as in individuals. The following three criteria are central areas for promoting reliability and, in a following step, for promoting the trustworthiness and acceptance of Artificial Intelligence in human-machine interaction (Trustworthy AI) (EU High-Level Expert Group 2019).

### Criterion 4: Quality of the available data

**Objectives:** Avoiding qualitatively insufficient data and the corresponding negative consequences; preventing distorted data sets, errors or misinterpretations, and discrimination; increasing the quality of statistical predictions by Artificial Intelligence; improving human-machine interaction through reliable data.

**Starting points:** Various points of reference can be identified for ensuring data quality. The only data to be collected are the data actually required for an application. It is important to develop clear ideas about the necessary data even during the development phase and to ensure that enough data can be generated for the intended purpose – also and especially in global competition (design). High-quality process data (not personal data) that has been collected can also be used for other purposes and in other contexts so that existing data can be used, and no additional data need to be collected (collateral benefit).

The quality of the data in terms of content enables more targeted design of reliable and precise human-technology interfaces. This helps to prevent errors and wrong conclusions as well as discrimination and the related negative consequences and to enhance the quality of the interaction results. For example, it is possible to avoid misinterpretations and increase data significance and thus data quality by applying adaptive confidence regions (ranges of expected values) to the data gathered. High data quality (e.g. with regard to consistency, comparability, reliability, validity concerning content) is also the basis for secure and adaptive systems that are trained to connect to human action, whereby the interaction can be "humanized" or oriented toward human interaction/communication and thus can be designed to be more humane (interaction quality). Conversely, this kind of more adaptive design also generates better data as a result of the interaction.

Software development can also contribute to data quality, for example by separating data structures, data processing, and data transport. This development principle ("separation of concerns") is important from a technical point of view, but also impacts other areas – such as safety or transparency.

### Criterion 5: Transparency, explainability, and consistency

**Objectives:** Implementing explainable Artificial Intelligence approaches; developing ways for making self-learning systems understandable and creating (graded) transparency about their decision-making processes and decisions; preventing demotivation of and excessive strain on workers through contradiction-free design of the human-machine interaction.

**Starting points:** In human-machine interaction, humans are often confronted with a system so complex that is incomprehensible to them. This can result in demotivation and rejection of such systems. In order to counteract this, self-learning systems must be designed so that they provide workers with basic information about their fundamental functionality, the purposes and objectives inscribed in them, their data focus and data, and the newly formed categories, "hypotheses," results, and above all conclusions and decisions or recommendations (output) that depend on the data focus and the data. This requires easily accessible solutions that are oriented toward the target group and that promote learning and experience. Depending on the task and role of the user and on the area of application, transparency can also be graded.

All information relevant to the interaction must be presented in a way that the people concerned can understand it. Approaches from the field of explainable Artificial Intelligence, which attempts to develop methods that are useful for making self-learning systems understandable, offer pointers to help solve this problem. They focus particularly on the system's ability of self-description and its conformity to expectations. Explainable Artificial Intelligence thus makes an important contribution to the development of responsible Artificial Intelligence, which is characterized by transparency, fairness, reliability, and orientation toward ethical concepts.

Another important aspect regarding cognitive and social strain on workers is consistent and contradiction-free design of the interaction with the self-learning system. For one thing, contradictory information and processes in direct interaction at the interface generate frustration and must be avoided. For another, special care must be taken that conflicting goals in enterprise strategy are not replicated or incorporated in the AI systems – such as frequently occurring contradictions and demands in the triangle of time spent, cost efficiency, and quality, or typical tensions between abstract planning (target situation) and concrete implementation (actual situation).

If these contradictions and tensions are inscribed in the technical work systems and are thus "objectified" rather than being socially mediated or negotiated, there is a risk that they may reduce the workers' scope for action and self-efficacy and thus bring about negative consequences of strain. All in all, users should only be faced with manageable situations, including in their interaction with Artificial Intelligence. As early as the design stage,

attention should therefore be paid to ensure that the interaction does not result in psychological distress and its consequences (which must be reviewed in the work process, for example by means of legally required risk assessments).

If self-learning systems in human-machine interaction are designed to be comprehensible, contradiction-free, appropriate, and as part of a responsible system, this forms a stable basis for workers to trust and ultimately accept them.

### Criterion 6: Responsibility, liability, and trust in the system

**Objectives:** Attributability of responsibility; competence and resources as well as control over the system as a prerequisite for assuming responsibility; extent of control over the system as an indicator of the type and extent of users' responsibility; starting points for design through trustworthy Artificial Intelligence concepts.

**Starting points:** Transparency (including objectives) and controllability of self-learning systems are necessary prerequisites if companies as well as workers are to assume responsibility. This also includes explicitly defining which information, resources, and competencies must be in place so that people can have agency in interacting with them. It is also important that the design of the human-machine interaction allows workers to act according to their work orientation and the existing rules and, if necessary, to terminate an interaction. Only then will users be willing and able to assume responsibility for their role in the interaction in the socio-technical system. Transparency and controllability are also required for self-learning systems to be perceived as trustworthy.

Estimating future system behavior is the basis for human control of the technical system. Human interaction with the system should therefore take place at an appropriate level of abstraction and be within the range of what can be expected. This scope of knowledge and action must be tailored to the interaction and must be determined and coordinated in advance (with as much participation as possible). Specifically, with respect to anchoring trust, responsibility, and liability, sufficient testing and technology assessment must be carried out in the system design phase to rule out undesired system behavior. It must be ensured that companies and workers are able to control the system, and the relevant responsibilities must be defined. The degree to which they control the system also essentially determines the type and extent of their responsibility. In practice, as well as in the basic legal sense, the ability to control is linked to the placement of responsibility within the legal entity. At the same time, this firmly emphasizes the experience of self-efficacy on the human's side.

In addition, the system design must allow for processes (input, processing, and output) to be tracked ex post and for the causes of any problems to be identified. The design must provide for appropriate methods capable of accomplishing this despite the high complexity.

Trust in a technology is an individual and socially constructed process in which the quality of trust and also the evaluation criteria can change on the basis of experience and knowledge. This requirement can be concretized even at the design stage by means of the aspects of comprehensibility, manageability, and meaningfulness of experiencing a situation as well as interaction with a (self-)learning system. The concept of trustworthy Artificial Intelligence, which was developed by the High-Level Expert Group on Artificial Intelligence established by the European Commission, provides important starting points and formulates a framework for the use of Artificial Intelligence based on ethical guidelines (EU High-Level Expert Group 2019).

## 2.3 Cluster 3: Reasonable division of work

The use of AI technologies and self-learning systems not only affects job profiles, competence requirements, and work processes, but also affects the "division of work" between human and machine in general. The aim is to find a reasonable "division of work" by means of human-centered design of the human-machine interaction and to provide workers with lasting relief and support in their work. At the same time, the focus is on questions of agency and situation control as well as topics concerning flexible and situation-specific adaptation of the systems. The following three criteria address the various requirements for a reasonable "division of work" between human and machine. Care should be taken that the details of the criteria are tailored to the tasks as well as to the qualification profile and competencies of the workers.

### Criterion 7: Appropriateness, relief from strain, and support

**Objectives:** The capabilities of humans and machines should complement each other to achieve a reasonable "division of work"; appropriate work content and requirements; lasting relief and support of workers through AI-based assistance systems.

**Starting points:** The use of self-learning systems in various industries and areas is creating a new distribution of tasks between humans and technology. The decisive factor here is that the different abilities and characteristics of humans and technology are already considered during the design of the interactive systems. Mutually complementing the specific strengths in question can create a mutually encouraging relationship, avoid negative consequences of strain, and achieve relief and support for workers. In this way, Artificial Intelligence also provides the opportunity to promote a positive workload while at the same time avoiding overload or excessive demands.

When developing and implementing interactive AI systems, it is important to ensure that the interaction is tailored to the qualifications and competences of the users – in terms of both the content and the form of interaction (e.g. reaction times or intensity of work). Even at the design stage, it is necessary that the design is specific to the situation and the

users: the technology should be adapted to humans. At the same time, workers must be empowered to "work together" with AI systems. The technical design itself should promote this (see Criterion 11). It is key not only for workers, but also for leadership and management, to be able to realistically assess the potentials, but also the limits of Artificial Intelligence and learning systems in order to use them in a meaningful way and to minimize errors of judgment.

In addition, human beings should not be weakened in their central role in the work process but strengthened by interacting with Artificial Intelligence. AI-based assistance systems should therefore be designed to relieve workers of (physically or psychologically) strenuous or dangerous activities. In addition, self-learning systems can support humans concerning complex matters and tasks – for example in difficult decision-making situations – and thus empower them.

### Criterion 8: Agency and situation control

**Objectives:** Targeted and transparent design of both agency and situation control in human-machine interaction; minimizing and avoiding risks and negative consequences of strain.

**Starting points:** A central demand of humane work design is that people should not be forced to orient their actions toward technical systems. However, part of the agency or situation control (Who initiates actions? Who coordinates a situation or sequence of interactions?) is inscribed in the technical system when humans interact with self-learning systems, and this is a typical characteristic of complex socio-technical arrangements. Interactive AI systems can be developed in such a way that they "ask" or "force" their users to take predetermined and sometimes even newly generated actions; as a result, the self-learning system itself takes the role of a kind of "actor" in its relationship to the persons in certain situations. Practical experience in the context of human-robot collaboration shows that self-efficacy is important to the workers on the shopfloor. In practical terms, this means that a high degree of agency and situation control can prevent dissatisfaction among workers.

If humans and technology are to collaborate and learn from each other in a highly interactive and complementary "division of work," then their roles and their agency must be clearly defined. It is important for it to be clear whether the AI application or the worker has agency at a particular point in time as well as when concrete transfers in the interaction occur. It must be transparent who is contributing what to a common process at any point in time, who has situation control for the sub-process in question, and where or how the sub-processes are connected to each other. It is necessary to define rules and to create ways to trigger transfers between human and machine (AI application) reciprocally in a targeted and transparent way – for example, if the system cannot handle a situation, or the human wants to intervene or requests support.

This is the starting point for potentially attributing responsibility, which can relieve the situation of the burden of unclear risks. In addition, it formulates the basis for transparent and comprehensible design of self-learning systems, opens up possibilities for human intervention and company regulations, and enables options for designing human-machine interaction in a way that promotes learning. Not least, clarifying agency transparently and in an interactive and cooperative process is the linchpin for a complementary "division of work" between humans and AI systems.

### Criterion 9: Adaptivity, error tolerance, and customizability

**Objectives:** Enabling self-learning systems to adapt to the users' needs and requirements and their work practice in a flexible and situation-specific way.

**Starting points:** It is hoped that self-learning systems will be able to adapt to changes in their environment and to deal with complexity more flexibly, meaning that they will be highly adaptive. If the human-machine interaction is to benefit the users long-term, the AI system must be designed to be highly socially adaptive.

This means that self-learning systems must not only be capable of translating requirements from their environment into their own system logic ("assimilative adaptivity"), but also of adapting their own processing logic to the needs of their environment and, above all, to the needs of workers ("complementary adaptivity").

This implies a large number of prerequisites for highly interactive AI systems in a work context. It means, for example, that their design must "plan" for AI systems to allow and even support unforeseen processes of appropriation during their use. This goes far beyond the usual requirements for technology development – such as robustness, error tolerance, and customizability.

Ultimately, this criterion refers to self-learning systems now being capable of enabling implicit programmability in the usage process. Thus, the design would no longer focus on the user experience, but on user empowerment.

## 2.4 Cluster 4: Supportive working conditions

Self-learning systems can take on human work and thus potentially reduce the opportunities for humans to perceive with their senses, to learn, and to practice, as well as to experience competence. When designing Artificial Intelligence, care should therefore be taken to address basic human needs in a targeted manner. This applies to task profiles and the "division of work" between human and machine as well as to the design of the user interface. In concrete terms, this means that the design of human-machine interaction must ensure scope for action and richness of work, enable humans and the AI system to learn from each other and gain experience, and be sensitive to the requirements and functions of communication, cooperation, and integration. Three further design criteria are presented below.

### Criterion 10: Scope for action and richness of work

**Objectives:** Safeguarding and (where appropriate) expanding the workers' scope of action (in particular autonomy and freedom to decide as well as a variety of options for action); inclusion of basic human needs for meaningful, motivating work that promotes health and personal development.

**Starting points:** Scope of action is understood to be the degree of autonomy and freedom to decide as well as the variety of opportunities available to a person at work to act and shape things. This applies to the goals, the work content and the concrete execution of tasks as well as to the structuring (organizational and technical) framework. In addition, the aspect of richness of work, i.e. the enrichment of activities with varied, challenging, and supportive content (job enrichment), should be taken into account. Expanding the scope for action and rich work are key aspects of the humanization of the realm of work and offer important points of reference for designing human-machine interaction.

Concerning the design and use of Artificial Intelligence, this means that care must be taken to ensure that these technical systems do not restrict the users' scope of action and do not take over precisely those parts of the work content that have a motivating, qualifying, and health-promoting effect. On the contrary, Artificial Intelligence should enable and expand the scope of action by making previously impossible actions achievable. The human-machine interaction when using Artificial Intelligence can itself provide new work content that is challenging and motivating and that promotes development.

### Criterion 11: Conducive to learning and gaining experience

**Objectives:** Designing human-machine interaction in a way that promotes learning and experience; facilitating humans learning from the machine and vice versa; comprehensible and adaptive design of systems for acquiring and integrating knowledge and experience; ensuring the transfer of data into information or of information into knowledge.

**Starting points:** Considering the essential differences between human and machine in acquiring, processing, storing, reproducing, retrieving, and applying knowledge is a prerequisite for implementing the criterion in order to meet the heterogeneous requirements both sides place on learning and to enable mutual learning.

Therefore, firstly, interaction with AI systems must be designed to promote users in learning and gaining experience. This concerns transparent (explainable AI) and mutually adaptive design in order to enable the acquisition of knowledge and experience in the usage process. Conflicting goals may emerge at this point, for example in relation to worker data privacy and individualized learning; such conflicts can, however, be solved by suitable regulations and corresponding design. This also concerns integrating specific learning content (qualifications) in human-machine interaction in a targeted manner as well as implementing design that takes this up in its didactic approach, for example design of assistance systems.

Secondly, designing the human-machine interaction to promote mutual learning can improve the AI system's performance and accuracy of fit by enabling users to interactively validate and, as appropriate, correct the learning content (data quality) and the learning behavior (links) of the intelligent system. Mutually learning-friendly design also increases the probability that people are willing to contribute their knowledge and experience to AI systems.

This type of complementary approach offers great opportunities, especially for dealing with complex situations. It is only in the socio-technical system that data can be contextualized to information and transformed into knowledge that can be applied to concrete situations and transferred. Only then can machine-learned content and human experience be integrated in a meaningful way.

Especially when AI systems take on farreaching and relevant activities and when opportunities for acquiring knowledge and experience may become scarce, designing human-machine interaction in a way that promotes learning and experience is of great importance – above all to maintain and expand knowledge, experience, and competencies and to strengthen users' self-efficacy – but also, of course, to promote innovation from within the processes and to enable users to assess the performance of the AI system and, as appropriate, to correct or improve it.

### Criterion 12: Communication, cooperation, and social embeddedness

**Objectives:** Dual sensitization of Artificial Intelligence for social contexts and structures; strengthening and supporting interpersonal communication, collaboration, and connectedness through AI systems.

**Starting points:** Communication, cooperation, and integration are essential basic conditions for high-quality and efficient work, sense-making, and integration. For one thing, this applies to the formal parts of work, which are planned and designed in advance and are reflected in formal work requirements, organization of work, and technology design. For another, it is about informal work practice, situational coordination, and innovations for flexibly coping with the parts of work that cannot be planned.

A major challenge for designing human-machine interaction lies in making Artificial Intelligence "sensitive" in two ways to social contexts and structures: for one thing, Artificial Intelligence can – to a very limited extent – act as a "cooperation partner"; for another, human-machine interaction must be designed in such a way that it does not prevent or replace necessary and beneficial interpersonal communication, cooperation, and connectedness, but at best even supports them in a goal-oriented manner.

AI-based technologies can support this in a variety of ways – for example, by relieving workers of standard communication, by independently collecting data from data silos and making them available for decision-making, by providing knowledge carriers with the information they need to make decisions, by "matching" or identifying knowledge carriers and connecting them in a situation-specific way, or by becoming "cooperation partners" themselves. AI-based technologies in interactive systems should be designed so that the technical system recognizes or takes up the human being's abilities and skills, i.e. competences, during the collaboration and that the functionalities in the interaction are adapted to this. For example, collaborative robots (cobots) working with people performing manual labor should be adapted to those individuals' physical abilities, skills, and experience (e.g. regarding process design).

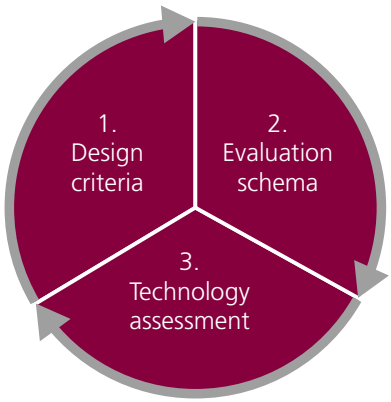
# 3. Implementing the criteria and outlook

The criteria presented are intended to provide an important impulse for human-centered and future-oriented design of human-machine interaction when using Artificial Intelligence long-term. In addition, they are to contribute to developing a new distribution of tasks between workers and technical systems – in this case, self-learning systems in particular.

The criteria are to be understood as guidance with concrete starting points for developing and deploying Artificial Intelligence (summary overview in Table 1). It should be noted that the design criteria must be concretized further for the application in question. In the process, conflicts of goals and interests emerge which must be negotiated and processed or regulated within the framework of existing law. In addition, the criteria can be interlinked with similar concepts (such as the approaches of human-centered design or experience-conducive technology design) and should be incorporated in ongoing activities – such as the development of norms and standards.

The set of criteria is one of three building blocks of a design tool (under development) for human-machine interaction employing Artificial Intelligence (Figure 2). The design criteria (1) provide a comprehensive overview of the design requirements. An evaluation schema (2) of the entire socio-technical arrangement is introduced as well in order to enable overall evaluation across these criteria from a different perspective. This involves evaluation of the quality and intensity of the interaction between humans and technology – following different levels of autonomy (acatech/Fachforum autonome Systeme 2017, Plattform Industrie 4.0 2019a) or different levels of criticality.

**Figure 2: Design concept for human-machine interaction**




Source: Huchler.

Beginning with the evaluation of human-machine interaction, one can extend the horizon of observation step by step by means of technology assessment to enable a comprehensive view of the consequences and requirements (3): from the potential effects on the concrete interaction situation of human-machine interaction on the ground, via the consequences for the working environment (such as cooperation, leadership, or organization of work) or the company (such as qualification, technology development and use of technology, product quality, or value creation concepts) to the consequences for society (such as jobs, income, education, cohesion, or consumer protection).

Not least, the design of human-machine interaction in highly interactive AI systems constitutes only a small part of the demands that Artificial Intelligence places on the transformation of work. Therefore, the design criteria should be supplemented and revised by a further white paper on questions of implementation and change management in companies in view of the increasing use of self-learning systems in the working environment.

**Table 1: Criteria for the design of human-machine interaction**

| **Protection of the individual** | |
|---|---|
| Protection of safety and health | • Avoiding risks to workers' physical and mental health<br>• Protecting against accidents and damage (personal injury and damage to property)<br>• Preventing negative physical or psychological consequences of strain |
| Data privacy and responsible performance monitoring | • Protecting personality rights, data economy, and purpose limitation of data use<br>• Avoiding data analysis for unjustified performance monitoring<br>• Developing a positive culture of performance feedback<br>• Transparency about data analysis and use; empowering workers to handle data transparency |
| Diversity sensitivity and non-discrimination | • Protecting against discrimination of individuals or groups<br>• Existing legal system as the basis for diversity sensitivity and non-discrimination |

| | |
|---|---|
| **Trustworthiness** | |
| Quality of the available data | • Avoiding qualitatively insufficient data<br>• Preventing distorted data sets, errors/misinterpretations, and discrimination<br>• Improving human-machine interaction through reliable data |
| Transparency, explainability and consistency | • Implementing explainable Artificial Intelligence approaches<br>• Developing methods useful for making self-learning systems understandable<br>• Creating (graded) transparency about decision-making processes of self-learning systems<br>• Preventing demotivation of workers through contradiction-free design of the human-machine interaction |
| Responsibility, liability and trust in the system | • Transparency and attributability of responsibility<br>• Competence and control over the system as a prerequisite for taking responsibility<br>• Extent of control over the system as an indicator of the type and extent of users' responsibility<br>• Starting points following the concepts of trustworthy Artificial Intelligence |
| **Reasonable division of work** | |
| Appropriateness, relief from strain, and support | • Appropriate work content and requirements<br>• Mutually complementing human and machine capabilities to achieve a reasonable division of work<br>• Lasting relief and support of workers through AI-based assistance systems<br>• Enabling workers to work with AI systems |
| Agency and situation control | • Targeted and transparent design of agency and situation control<br>• Minimizing and avoiding risks and negative consequences of strain |
| Adaptivity, error tolerance, and customizability | • Enabling self-learning systems to adapt flexibly and situationally to the needs and requirements and to the working practice of the users |
| **Supportive working conditions** | |
| Scope for action and richness of work | • Securing and, as appropriate, expanding workers' scope for action (especially autonomy and freedom to decide as well as the variety of possible actions)<br>• Inclusion of basic human needs for meaningful, motivating work that promotes health and personality development |
| Conducive to learning and gaining experience | • Facilitating humans learning from the machine and vice versa<br>• Comprehensible and adaptive design of systems for integrating knowledge and experience<br>• Ensuring the transfer of data into information or of information into knowledge |
| Communication, cooperation, and social embeddedness | • Dual sensitization of Artificial Intelligence for social contexts and structures<br>• Supporting interpersonal communication, collaboration, and connectedness<br>• Artificial Intelligence as a cooperation partner |

# About this White Paper

This paper was prepared by the Working Group Future of Work and Human-Machine Interaction of Plattform Lernende Systeme. As one of seven working groups, it examines the potentials and challenges arising from the use of Artificial Intelligence in the realm of work and the lifeworld. It focuses on questions of transformation and the development of humane working conditions. In addition, it addresses the requirements and options for qualification and lifelong learning as well as starting points for designing human-machine interaction and the division of work between humans and technology.

**Authors:**
**Dr. Norbert Huchler** (Lead author), Institut für Sozialwissenschaftliche Forschung e. V. (ISF-München)
**Prof. Dr. Lars Adolph**, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)
**Prof. Dr. Elisabeth André**, Universität Augsburg
**Prof. Dr.-Ing. Prof. e. h. Wilhelm Bauer**, Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO und Universität Stuttgart
**Nadine Bender**, KUKA Deutschland GmbH
**Dr. Nadine Müller**, Vereinte Dienstleistungsgewerkschaft (ver.di)
**Dr. Rahild Neuburger**, Ludwig-Maximilians-Universität München
**Dr.-Ing. Matthias Peissner**, Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO)
**Prof. Dr. Jochen Steil**, Technische Universität Braunschweig
**Prof. Dr. Ing. Sascha Stowasser**, Institut für angewandte Arbeitswissenschaft (ifaa)
**Oliver Suchy**, Deutscher Gewerkschaftsbund (DGB)

**The Working Group is directed by:**
**Prof. Dr. Elisabeth André**, Universität Augsburg
**Prof. Dr.-Ing. Prof. e. h. Wilhelm Bauer**, Fraunhofer-Institut für Arbeitswirtschaft und Organisation IAO und Universität Stuttgart

**The members of the Working Group:**
**Prof. Dr. Lars Adolph**, Bundesanstalt für Arbeitsschutz und Arbeitsmedizin (BAuA)
**Prof. Dr.-Ing. Jan C. Aurich**, Technische Universität Kaiserslautern
**Vanessa Barth**, IG Metall
**Klaus Bauer**, TRUMPF Werkzeugmaschinen GmbH + Co. KG
**Nadine Bender**, KUKA Deutschland GmbH
**Prof. Dr. Angelika Bullinger-Hoffmann**, Technische Universität Chemnitz
**Prof. Dr.-Ing. Barbara Deml**, Karlsruher Institut für Technologie (KIT)
**Prof. Dr. Prof. h.c. Andreas Dengel**, Technische Universität Kaiserslautern und Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI) GmbH
**Dr. Jan-Henning Fabian**, ABB AG
**Prof. Dr.-Ing. Sami Haddadin**, Munich School of Robotics and Machine Intelligence, Technische Universität München

**Prof. Dr. Michael Heister**, Bundesinstitut für Berufsbildung (BIBB)
**Prof. Dr.-Ing. Rolf Hiersemann**, Hiersemann Prozessautomation GmbH
**Dr. Norbert Huchler**, Institut für Sozialwissenschaftliche Forschung e. V. (ISF-München)
**Dr. Nadine Müller**, Vereinte Dienstleistungsgewerkschaft (ver.di)
**Dr. Rahild Neuburger**, Ludwig-Maximilians-Universität München
**Dr.-Ing. Matthias Peissner**, Fraunhofer-Institut für Arbeitswirtschaft und Organisation (IAO)
**Prof. Dr.-Ing. Annika Raatz**, Leibniz Universität Hannover
**Prof. Dr.-Ing. Jürgen Roßmann**, Rheinisch-Westfälische Technische Hochschule Aachen
**Prof. Dr. Christoph M. Schmidt**, RWI – Leibniz-Institut für Wirtschaftsforschung und Ruhr-Universität Bochum
**Dr. Anke Soemer**, Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e. V.
**Prof. Dr. Jochen Steil**, Technische Universität Braunschweig
**Andrea Stich**, Infineon Technologies AG
**Oliver Suchy**, Deutscher Gewerkschaftsbund (DGB)
**Prof. Dr.-Ing. Sascha Stowasser**, Institut für angewandte Arbeitswissenschaft (ifaa)
**Dr. Hans-Jörg Vögel**, BMW Group
**Jochen Werne**, Prosegur Cash Services Germany GmbH

**The Working Group is supported by:**
**Dr. Chi-Tai Dang**, Universität Augsburg
**Dr.-Ing. Jan Harder**, Munich School of Robotics and Machine Intelligence, Technische Universität München
**Dr. Andreas Heindl**, Managing Office, Plattform Lernende Systeme
**Dr.-Ing. Michael Wächter**, Technische Universität Chemnitz

**Editorial**
**Dr. Andreas Heindl**, Managing Office, Plattform Lernende Systeme
**Dr. Ursula Ohliger**, Managing Office, Plattform Lernende Systeme
**Alexander Mihatsch**, Managing Office, Plattform Lernende Systeme

## About the Plattform Lernende Systeme

Designing self-learning systems in the interests of society – this was the aim of the Plattform Lernende Systeme, which was initiated in 2017 by the Federal Ministry of Education and Research at the suggestion of the Autonomous Systems Forum of the High-Tech Forum and acatech – The National Academy of Science and Engineering. The platform bundles the existing expertise in the field of Artificial Intelligence and supports Germany's further path to becoming a leading international technology provider. The approximately 200 members of the platform are organized in working groups and a steering committee. They demonstrate the personal, social, and economic benefits of self-learning systems and identify challenges and design options.

# References

**acatech (2016)**: Neue autoMobilität – Automatisierter Straßenverkehr der Zukunft. Online unter: https://www.acatech.de/publikation/neue-automobilitaet-automatisierter-strassenverkehr-der-zukunft/ (accessed: 20 March 2020).

**acatech/Fachforum Autonome Systeme (2017)**: Chancen und Risiken für Wirtschaft, Wissenschaft und Gesellschaft (Abschlussbericht). https://www.acatech.de/publikation/fachforum-autonome-systeme-chancen-und-risiken-fuer-wirtschaft-wissenschaft-und-gesellschaft-abschlussbericht (accessed: 20 March 2020).

**EU High-Level Expert Group (2019)**: Ethics Guidelines for Trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai (accessed: 20 March 2020).

**Huchler, Norbert (2019)**: Assimilierte vs. Komplementäre Adaptivität. Grenzen teil-autonomer Systeme, in: Hirsch-Kreinsen, Hartmut/Karačić, Anemari (Hrsg.): Autonome Systeme und Arbeit. Perspektiven, Herausforderungen und Grenzen der Künstlichen Intelligenz in der Arbeitswelt, Bielefeld, pp. 139–180.

**Huchler, Norbert (2016)**: Die Grenzen der Digitalisierung. Neubestimmung der hybriden Handlungsträgerschaft zwischen Mensch und Technik und Implikationen für eine humane Technikgestaltung. In: „Digitalisierung, IT und Arbeit" HMD Praxis der Wirtschaftsinformatik, issue 53 (1), Wiesbaden: Springer, pp. 109–123.

**Manzeschke, Arne/Weber, Karsten/Rother, Elisabeth/Fangerau, Heiner (2013)**: Ethische Fragen im Bereich altersgerechter Assistenzsysteme. Ergebnisse der Studie. https://www.technik-zum-menschen-bringen.de/service/publikationen/ethische-fragen-im-bereich-altersgerechter-assistenzsysteme (letzter Zugriff: 20.03.2020).

**Plattform Industrie 4.0 (2019a)**: Industrie 4.0 gestalten. Souverän. Interoperabel. Nachhaltig. Fortschrittsbericht 2019. https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/hm-2019-fortschrittsbericht.html (accessed: 20 March 2020).

**Plattform Industrie 4.0 (2019b)**: Technologieszenario Künstliche Intelligenz in der Industrie 4.0. https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/KI-industrie-40.html (accessed: 20 March 2020).

**Plattform Lernende Systeme (2019)**: Künstliche Intelligenz und Diskriminierung: Herausforderungen und Lösungsansätze, Whitepaper der AG3 – IT-Sicherheit, Privacy, Recht und Ethik. https://www.plattform-lernende-systeme.de/publikationen-details/kuenstliche-intelligenz-und-diskriminierung-herausforderungen-und-loesungsansaetze.html (accessed: 20 March 2020).

# Imprint

**Editor**
Lernende Systeme –
Germany's Platform for Artificial Intelligence |
Managing Office | c/o acatech
Karolinenplatz 4 | D-80333 Munich
www.plattform-lernende-systeme.de

**Design and Production**
PRpetuum GmbH, Munich

**Status**
June 2020

**Image credit**
Westend61/gettyimages/Titel
VICTOR/iStock/S. 9

In case of questions or comments regarding this publication please contact Johannes Winter (Director of the Managing Office):
kontakt@plattform-lernende-systeme.de

Follow us on Twitter: @LernendeSysteme

**Recommended citation**
Norbert Huchler et al. (eds.): Criteria for Human-Machine Interaction When Using AI – Approaches to its humane design in the realm of work. White paper from Plattform Lernende Systeme, Munich 2020.