Deutscher Ethikrat

Humans and Machines – Challenges of Artificial Intelligence

OPINION · EXECUTIVE SUMMARY & RECOMMENDATIONS

20 March 2023

The full text of the Opinion “Humans and Machines – Challenges of Artificial Intelligence” as well as all accompanying information and documentation issued by the German Ethics Council on the topic are available at https://www.ethikrat.org/en/publications/opinions/humans-and-machines.

Published by the German Ethics Council

Jägerstraße 22/23 · D-10117 Berlin
Phone: +49/30/20370-242 · Fax: +49/30/20370-252
Email: kontakt@ethikrat.org
www.ethikrat.org

Title of the original German edition: Mensch und Maschine – Herausforderungen durch Künstliche Intelligenz

English translation: Noah Harley
Layout: Torsten Kulick
Titelillustration: pinkeyes/Shutterstock.com

>> CONTENTS

Introduction

1) Digital technologies and artificial intelligence (AI) have found their way into nearly every corner of contemporary life, both public and private. In considering the broader social implications of these developments from an ethical standpoint, it is not enough to understand the technologies involved. The ways in which people interact with them must also be taken into account, whether they use them directly or are impacted by their use. One key question considers the implications of delegating tasks previously reserved for humans to machines: Will the use of AI increase or diminish human authorship and options for action?

2) In pursuing this question, the German Ethics Council continues to explore a set of topics it has previously addressed in its Opinions "Big Data and Health – Data Sovereignty as the Shaping of Informational Freedom" (2017) and "Robotics for Good Care" (2020). The present

Opinion also takes up a request from the President of the German Bundestag in October 2020 to draft a multidisciplinary treatise on the ethical issues posed by human–machine relations.

3) The Opinion is divided into three sections. The first addresses the *underlying technical and philosophical* aspects of the topic. The second section follows with concrete examples, analysing the ethics of AI as it appears in four select fields: *medicine*, *education in schools*, *public communication and opinion formation*, and *public administration*. The final section identifies ten *overarching topics* relevant to each of these four fields, including general recommendations.

>> PART I: TECHNICAL AND PHILOSOPHICAL FOUNDATIONS

Major developments and technical foundations of artificial intelligence

4) The idea of machines whose capabilities match or even surpass our own in definitive human pursuits like cognition, learning or action can be traced back to Greek mythology, millennia before the invention of software systems. It was only with the development of the first computers in the twentieth century, however, that machine intelligence became a tangible reality. In 1950, mathematician Alan Turing formulated a criterion for determining AI, in what later came to be known as the Turing Test, whereby machine intelligence is said to be present when a human observer is not able to distinguish a machine's behaviour from that of a person.

5) Early research into AI operated on the premise that human learning or intelligence could be described precisely enough to enable a machine to simulate it. Examples of early research topics that maintain their relevance today include pattern recognition, language processing,

the capacity for abstraction, creativity and flexible problem-solving. Improvements to computer hardware and programming languages soon gave rise to a great sense of optimism surrounding the potential of machine intelligence. The following decades saw successive waves of enthusiasm followed by "AI winters", where disappointment at a perceived lack of practical results prevailed and funding was cut back.

6) AI research in the late twentieth century was marked by continued progress on several key fronts, the rise of parallel data processing methods and the Internet, and growing involvement on the part of research organisations, the military and industry. A critical discourse also arose in tandem to these developments, with computer ethics establishing itself as an independent discipline that raised increasing philosophical doubt as to whether the specific vision some researchers laid out for the emergence of general or strong AI could – or should – ever become reality.

7) The appearance of three trends around the turn of the millennium gave AI a dynamism that continues into the present day: First, a dramatic rise in computer processing power and miniaturisation; second, increasingly tight networks between digital systems; and with those, third, new possibilities for data collection and analysis.

8) The result has been computer introduction into many aspects of daily life, including the countless commonplace objects like cellphones, clocks and household devices that are now "smart", i.e. networked and equipped with sensors. These devices and the data networks that link them have, in turn, generated socio-technological data ecosystems that give an increasingly accurate and comprehensive digital picture of our movements, actions, characteristics and preferences. Digital representations of this sort not only make data analysis possible, but can also impact human behaviour directly, with people receiving information or recommended actions on their basis.

9) Throughout, data and the metadata that accompany it – both of which can vary drastically in type and quality – serve as the foundation for these kinds of digital operations and interactions. The quality of a given data set does not simply depend on how accurate, complete, current or detailed the data are. It also depends on the relationship between the context from which the data were originally taken and the one in which they will be used. Data can be more or less suitable for a given question or set task. Failing to address quality or suitability concerns in a timely or sufficient manner can result in errors, bias and misleading analysis.

10) The hardware and infrastructure available for managing and using data also play a decisive role in how well data-driven applications will perform. At present, these include services accessible via the Internet (cloud computing) and powered by large-scale facilities that specialise in storing and/or analysing data, as well as increasingly powerful options for processing data locally, at least in part, on the devices collecting it (edge computing).

11) At the heart of any form of data processing are algorithms, processing instructions that prescribe how entered data should be handled until the desired output value is reached, generally using clearly defined step-by-step rules. Of particular importance in current AI research are statistical analyses that recognise regularities in data and identify connections between individual characteristics, allowing for predictions about similar data sets or future developments. If the aim is to demonstrate a cause-and-effect relationship between properties, further analysis and investigation are usually necessary to offer an explanation that is at once plausible and empirically verifiable – by way of experiments, for example.

12) Statistical analyses often contain ambiguities that cannot be fully resolved. What is more, minimising certain sources of errors can amplify others. This means that the most important errors to account

for in statistical analysis will always depend on the specific question or purpose at hand, and that in many cases it will not simply be a question of technology or methods but also of ethics.

13) The algorithmic methods and systems used in AI are often grouped together under the catchphrase "machine learning", and distinguished by their ability to use data to optimise a wide range of functions including pattern identification and model construction. To do so, an algorithm undergoes an initial training phase, during which it first constructs and then gradually refines a model for pattern recognition by repeatedly analysing training data sets.

14) Machine learning encompasses a range of approaches. In supervised learning, the relationship between the input data and the desired output is already fixed in the training data set. One example would be a set containing images of healthy skin and instances of skin cancer, with a label for each image specifying the category to which it should be assigned. In unsupervised learning, however, the algorithm "searches" for patterns in the data on its own, without the training data being labeled ahead of time. In reinforcement learning, the algorithm optimises its activity for particular goals, receiving feedback after each cycle of the training phase as to whether the last attempt brought the system closer to or further away from its goal.

15) Deep learning is a subset of machine learning that is especially well-suited to handling larger quantities of data. It has been an important driver for a host of AI applications in recent years. The approach makes use of "neural networks", which are loosely based on connective structures in the brain.

16) As a general rule, the strategies an algorithm ultimately derives for dealing with a set task in training are not fully comprehensible even to trained personnel with complete access to the algorithm's code (blackbox). While possibilities do exist for making algorithmic

processes sufficiently transparent, interpretable or explainable to a given target group (explainable AI), selecting and applying them poses a technical challenge.

17) The combined effects of developments in computer hardware and software, networking and data production have generated a world of potential uses for AI. AI systems are now able to beat humans in demanding strategy games like chess or Go (MuZero). They produce complex texts whose computer generated origin is often unrecognisable (ChatGPT).

18) In the present Opinion, the German Ethics Council considers four fields of activity in which AI has either already brought about especially broad changes or may do so in the near future. In *medicine*, for example, machine learning holds out the prospect of improved diagnostics and individualised recommendations for preventive care and treatment. *School education* meanwhile is seeing a wide array of AI-supported methods emerge for more effectively transmitting knowledge and skill sets. A large part of the information exchanged in the field of *public communication and opinion formation* already runs through digital or social media platforms that rely on algorithms. Finally, the algorithmic systems used to assist with decisions and prognoses in *public administration* affect the lives of many, for example in the assessment or monitoring of individuals in the welfare or police sector.

19) A regulatory landscape has started to take shape in response to the challenges that these shifts pose for human interaction. This consists on the one hand of a profusion of guidelines ranging from codes for individual companies and directives for professional associations to mechanisms at the national or international level. On the other hand, the legal framework also continues to develop, for example, the media laws passed in Germany. Socio-technological developments in the field of AI are often driven by IT companies with an active

international presence, giving greater significance to transnational regulation. Examples in the EU include the General Data Protection Regulation and the proposed Artificial Intelligence Act.

20) Taking these developments into account, the German Ethics Council does not focus on the legal landscape in this Opinion. Instead, it considers what implication changes in the digital world may hold for human coexistence by way of philosophical inquiry into basic anthropological terms that lie at the heart of our self-conception as humans. Building on this, it develops an approach to human-technology relations in which it is crucial to understand how the delegation of human activities to machines and algorithmic systems affects central anthropological concepts and, in particular, expands or diminishes human authorship.

Central concepts and philosophical foundations

21) The meaning of the term artificial intelligence has changed over the years, varying both within and among different professions and disciplines. Within the definition, an important distinction exists between weak and strong; the latter designates a form that is either comparable with or even superior to human ability. Special versus general AI and narrow versus broad AI are two other conceptual pairs that look to describe forms or degrees of AI bordering on human intelligence.

22) The distinctions between special, narrow or weak AI on the one hand and general, broad or strong AI on the other do more than simply describe two opposite poles. Rather, underlying each – particularly the distinction between weak and strong AI – is a particular understanding of intelligence and a different answer to a key question: Are the differences that exist between human and artificial intelligence qualitative and categorical or simply a matter of degrees that can in principle be overcome?

23) One important factor in this context lies in the relative breadth or narrowness of what AI is capable of. Most AI applications perform their assigned tasks within clearly established, narrowly defined parameters or fields. There is also the matter of whether intelligence is linked to certain mental preconditions that are doing more than simply *simulating* understanding. The question thus arises as to whether machines can ever fully possess intelligence in a general or strong sense or whether properties that are specific to humans form necessary prerequisites instead.

24) Responses vary according to the theoretical anthropological model in use. Most behaviourists, for example, would see a humanoid robot with a flawless range of motion and facial expressions and gestures similar to those of humans as an instance of broad or even strong AI, provided it was capable of faithfully simulating human cognitive capacities. Other theoretical models would dispute that a form of strong AI were present, as even a perfect simulation would not guarantee the robot's ability to exhibit mental states, show insight and judgement or access emotions like hope or fear.

25) The present Opinion works from the premise that while the difference between narrow and broad AI is of a quantitative or gradual nature, the emergence of strong AI would signal a qualitative leap. *Narrow AI* refers to applications which simulate human capacities in a given domain in order to carry out specific tasks. *Broad AI* expands an AI's range of potential applications beyond individual domains. The term *strong AI* designates a type of artificial intelligence that goes beyond even a perfect simulation of human cognition to exhibit mental states and the capacity for insight and emotions.

26) Notions of human intelligence provide one important basis for discussions surrounding the current and future potential of AI. From a psychological perspective, intelligence should be viewed as a hypothetical construct. While it can be sketched verbally, i.e. with

reference to concepts such as understanding, judgement, deduction, goal-oriented action, rational thought or engaging effectively with one's environment, it is not directly observable. Intelligence tests offer one form of operationalisation as they provide humans with a context in which to demonstrate behaviour that can then be deemed more or less "intelligent" according to a pre-existing theoretical definition.

27) Whether intelligence is a single, uniform capacity or comprises multiple capacities that may potentially operate independently cannot be conclusively answered empirically. The link between intelligence and creativity is another contentious topic of relevance for AI. A key distinction here lies between convergent thinking, which proceeds by way of logical conclusion to reach a single or optimal solution, and divergent thinking, a hallmark of creativity that is able to find multiple alternative solutions for a given set of requirements.

28) The concept of intelligence has steadily expanded in recent years, now encompassing terms such as social or emotional intelligence. At the same time, a field of research has coalesced around the terms embodied, embedded, enactive and extended cognition. It considers the role the body and environment play in intelligence and cognitive ability through the lenses of philosophy, psychology and robotics. This conceptual broadening has prompted fundamental questions, if they had not already arisen, about how to apply the concept of intelligence to technical artefacts. The word "intelligence" as it appears in the term "artificial intelligence" should, therefore, rather be taken as a metaphor whose descriptive and explanatory power stands in need of greater clarification.

29) Long before the term intelligence came into vogue, reason was used as a term to describe humans' special capacity for orienting themselves in the world and taking responsibility for their actions, bringing coherent structure to their lives in the process. Intelligence is an

important prerequisite for reason, though not a sufficient condition in and of itself.

30) Reason is a highly complex term, comprising a multidimensional set of relationships between ways of thinking, reflecting and operating that are interwoven into a complex social and cultural fabric, and which in their entirety aim at apprehending reality as effectively as possible. A fundamental distinction in this case lies between theoretical reason, which looks to acquire knowledge so as to arrive at valid empirical or a priori judgements, and practical reason, which seeks a coherent, responsible course of action that will allow a person to lead a fulfilling life.

31) Particularly in the case of theoretical reason, a number of parallels to the working methods of AI systems seem to present themselves: the capacities of processing information, learning, drawing logical deductions, following rules consistently and establishing meaningful connections between stored information figure centrally in both. Yet critical differences arise upon closer inspection. Not only does human memory operate differently in multiple respects from a computer's storage apparatus; there is no technical substitute for the human practice of judgement. The AI systems that exist to date, at least, do not possess the interpretive understanding, intentionality or reference to a reality beyond language necessary to do so.

32) This is clearer still in the case of practical reason, which is even more complex, aiming not only at well-founded individual practical judgements but the most fitting and responsible course of action for the long term so as to give a person's actions coherent shape and allow them to lead a fulfilling life. Doing so requires multiple individual components; whether or not technical artefacts will be able to simulate them is the subject of fierce debate.

33) Among these components are, first, an understanding of important terms in the human moral vocabulary for designating morally relevant aspects, values and positions; second, a capacity to distinguish and feel empathy; third, an ability to weigh conflicting aspects and values; fourth, the adoption of a reflective approach to rules of varying degree and scope; fifth, the ability to intuitively grasp complex tasks and circumstances; sixth, a capacity for discernment; seventh, the ability to justify moral judgements and resulting actions; and eighth, affect and impulse control, so that the practical judgements reached in a given case can be translated into action.

34) While partial overlap is entirely possible between the range of modern AI systems' capabilities and the complex phenomenon of human reason, it is critical to keep in mind that the capacities listed above do not exist independently of one another but must be conceived as forming a rich web of interactions, mutual feedback and conditioned relationships. They constitute an integral part of human nature in all its complexity, which must be apprehended as an indivisible, body–soul entity. Human reason should always be seen as embodied reason. Nor can practical reason be viewed purely from an individual perspective. Every individual belongs to a social and cultural environment with a lasting impact on their socialisation, meaning that supraindividual cultural factors must also be taken into account when interpreting practical reason.

35) Any sufficient understanding of the matter, especially where the use of practical reason is concerned, is tightly bound up in our baseline understanding of ourselves as individuals capable of action. Not every human activity that has an effect on the immediate environment should be seen as action per se; rather only those activities that are purposive, intentional and controlled. Assuming that machines do not operate purposively, i.e. have no intentions of their own, makes it paradoxical to attribute "actions" to them in this narrow sense.

36) Even so, since the turn of the millennium the discourse around AI has increasingly asked whether there are contexts in which machines could, in fact, be said to act in a broader sense, outside the narrower definition given above – in the event that decisions are fully delegated to software systems, for example. Linked to this question is a debate about whether and to what extent systems with growing autonomy, i.e. those that operate without human involvement, can subsequently be held responsible as the "agents" of their "actions", where questions of liability are concerned for instance.

37) Even if machines do execute complex operations that change the world and show themselves capable of dealing flexibly with major challenges to humankind, they are not accomplishing those changes intentionally. This means further that they cannot be held responsible for them in a moral or legal sense. Against this backdrop, it seems sensible to reserve the term "action" in the restricted sense for humans, both to avoid inflating the concept of what an actor is and to allow for conceptual boundaries to be drawn.

38) The concept of agency or authorship plays a decisive role in this context, connoting the universal human experience of viewing oneself and others as the originators of certain events or conditions. The capacity for authoring our actions may be regarded as a basis for autonomy or the fact that we as humans are able to guide our actions by maxims we ourselves have set.

39) The circumstances and results surrounding a given action may prove consequential when evaluating it from a moral or legal standpoint. In addition to any intended results, for example, a given action may also bring results that are unintended but still discernible to the actor. This, in turn, holds implications for the concept of negligence, which is an important issue in the field of AI. And even if it is primarily individuals who act, this does not rule out a concept of collective action

where multiple people have acted in a coordinated fashion from the outset.

40) Technology can also exert a considerable influence on how humans act or experience action. The steady saturation of our lived environment with machines that feature increasingly powerful information technology has resulted in hybrid socio-technological scenarios where humans and machines are tightly entwined and interact in complex ways. Moreover, many systems are now able to imitate certain human activities so effectively that their simulation passes for intentional human acts. This makes it prudent to stick to a narrow definition for action bound to the central criterion of intentionality.

41) Intentionality is an equally important criterion for assigning responsibility within the human–machine interactions that play out across ever more complex socio-technological networks. The concept of responsibility can be grasped in terms of a fivefold relationship: Who (the responsible subject) is responsible to whom (the affected party) before whom (authority) for what (object of responsibility), and under what norm?

42) Conversations around responsibility when it comes to scientific and technical progress should bear in mind that evaluating the practical consequences of new developments often occurs under a high degree of uncertainty that cannot be fully eliminated. Attribution of responsibility must, therefore, take into account the dimension of acting under uncertainty.

43) Moral responsibility can only be assumed by natural persons who possess the capacity to act, i.e. who are able to influence, and thus bring about changes to their environment in an active, purposeful and controlled manner. Were the same to apply to machines, they would also be capable of bearing responsibility. In that case one would have to assign them the status of personhood, although given the qualitative

developments anticipated for machine systems this would not be appropriate either at present or in the foreseeable future. Responsibility cannot, therefore, be assumed directly by automated systems but only by the humans standing behind these systems in different roles, where applicable in the context of institutional responsibility.

44) It can often be difficult to determine who exactly bears what degree of responsibility in each case. The multifaceted ties of responsibility linking individuals, institutions and the state become even more complex when their interactions are even partially supported or facilitated by algorithmic systems, whose operations may at times be anything but transparent or appear autonomous. Against such a backdrop, it is essential to appropriately configure multi-actor responsibility.

45) Action, reason and responsibility present central concerns in humanist philosophy. Humans are equipped with agency, and thus with authorship over their own lives. Their freedom gives them responsibility for shaping their actions. Freedom and responsibility are two mutually-dependent aspects of human authorship. Authorship, in turn, is tied to a capacity for reason.

46) The phenomenon of being affected by reasons is at the centre of this trinity of reason, freedom and responsibility. Practical reasons argue for actions, while theoretical reasons argue for convictions. As a general rule, reasons for doing one thing over another must be weighed against each other. Conflicting reasons give rise to a process of deliberation, which must then be systematised in terms of an ethical theory.

47) Human life is characterised by reactive attitudes and moral sentiments, which are accompanied by normative reasons. Freedom comes into play to the extent that we set these attitudes and sentiments aside when we learn that a person was not free in their actions. This practice of ascribing freedom and responsibility is essential for

the foundation of moral judgement. The norms of morality and law are baseless without assuming a sense of responsibility for humankind, and thus its capacity for freedom and reason.

48) The neurosciences offer one challenge to this humanist perspective: Empirical studies showing, for example, that the motor centre of the brain begins preparing for movement before one has consciously decided to move will intermittently be cited as evidence that freedom, and with it human responsibility does not exist. In fact, such findings are open to different interpretations and are not a suitable way of refuting human freedom and responsibility.

49) A second critique of humanist philosophy draws its inspiration from debates around AI. It flits between transhumanism on the one hand, which sets the goal of transcending human limitations and opening up new dimensions for human potential through human–machine symbioses, and a mechanistic paradigm that reduces the human psyche to the model of an algorithmic system on the other. This latter element is particularly relevant for this Opinion, as it exerts considerable influence on how interactions between people and machines are conceived, as well as their subsequent impact on human self-understanding.

50) Mechanistic paradigms view humans materialistically as machines or alternatively interpret machines as animate, possessing mental states and operating on the same plane as humans. The sometimes widespread tendency in AI discourses to equate an external indistinguishability of human and machine performance with the assumption of intelligence and capacity for thought of such machines is the result of certain theoretical preconceptions, especially of a behaviourist and functionalist nature.

51) Behaviourism looks to explain human behaviour on the basis of stimulus–response schemata that can be described with great precision,

and thus to transform psychology into an exact science. In so doing, the inner lives of organisms are completely overlooked. Functionalism rests on the assumption that mental states can be completely understood in terms of their function, and that questions regarding their mode of being can and should be set aside in place of a precise description of their function. Multiple realisability – the thesis that certain mental events, properties or states can come about through entirely different physical events, properties or states – also seems to make it possible to ascribe mental states to computers, even though they do not possess any biological structures.

52) Critiques of functionalism will make reference to phenomenal consciousness, according to which the mental states of a being crucially depend on qualities of sensation that do not disclose themselves through external behaviour alone. This type of consciousness places certain limitations on the possibility of judging another living being's quality of experience or mental states, and makes functionalism's human–computer analogy seem a dubious reduction.

53) A further argument against functionalism comes from philosopher John Searles' "Chinese Room" thought experiment, in which a person responds to a series of questions in Chinese from a sealed room using a precise set of instructions. It is not this person who knows the Chinese language, nor a translation computer, but those who created the instructions or the algorithm for answering the questions.

54) Counterarguments to functionalist machine paradigms illuminate the importance of general life experience for reason. Human reason is embodied reasoning. The body is both the point of departure and an essential component for any perception or sensation, as well as the precondition for living a human existence in the world and developing relationships to others. This means that the emergence and execution of our cognitive abilities are linked to our sensory world and corporeality, and our existence as social and cultural beings.

55) This also limits the extent to which human reason can be formalised and simulated. Acquiring human experience is always bound up in an interpretive process and presupposes some form of involvement or engagement. Here, too, the body has an important role to play. It allows action that would not be possible simply by means of conscious planning and calculation. This, in turn, substantiates why it is not possible to simulate thinking per se, implying a limit as to how far AI can be developed.

56) Based on the above considerations, it can be summarised that human intelligence is indissolubly linked to the manifold dimensions of the human lifeworld. It operates guided by reasons and an is expression of accepted values and norms. It is questionable whether this type of practice – led by reasons, conditioned by multiple dimensions, embedded in a sociocultural context, and coherent in itself – could ever be conceivable even for complex machine systems.

Human–technology relations

57) Humans develop, design and employ technology as a means to an end. Not infrequently however, delegating human tasks to machines in a more or less comprehensive manner – up to and including replacing human actors completely – subsequently impacts human options for action, skill sets, authorship and assumption of responsibility, either expanding or diminishing them. The three terms of *expansion*, *diminishment* and *replacement* serve as an analytical matrix in this Opinion.

58) Under the theory of social constructivism, technology design follows ends typically set by humans and shaped by the respective priorities of a society. Technological determinism, on the other hand, views internal dynamics as the defining factor, especially those set by economic relations, to which it contends humans and societies ultimately must submit and adapt. In reality, both processes are at play

simultaneously. The relationship between humans and technology rests on a dynamic of cocreation that might also be described as co-evolution. Social contexts and normative criteria on the one hand, and technologies on the other develop in tandem through mutual interaction.

59) Taken as a whole, technology can become a sort of second nature, setting the baseline conditions and terms of success for the continuation of human life, and shaping how we view the world and solve problems. New technology is thus often itself the product of humans approaching and relating to the world through one kind of technology or the other. The increasingly complex relationship between humans and technology or humans and machines also changes how those relationships are perceived. In systems guided by AI, what previously appeared to be clear distinctions between humans and technology are now less so. In our everyday speech as well, the anthropomorphising of digital technologies is well advanced, for example in the attribution of abilities such as thinking, learning, making decisions or showing emotion to AI and robots.

60) Subject–object relations between humans and technology are likewise changing; within networked systems, people will at times play the role of subject, but at others the role of object. In cases where software systems are given charge of decisions over people, for instance when assigning social services, the people become the objects of the system's "decisions", while the system acts as if it were the subject.

61) Different approaches look to describe these developments in terms of multi-level interactions between human and technology.

62) In each of these approaches, responsibility continues to rest with people. Using AI systems can nevertheless have morally problematic consequences that subsequently affect peoples' actions. Human action is,

therefore, neither completely autonomous nor completely socially or technically determined, but increasingly socio-technically situated.

63) In many cases, AI has distinctly positive effects in the sense of expanding the possibilities of human authorship. Yet as technology and innovations diffuse into society, are used and become more and more tightly bound up in our everyday routines, it can also diminish opportunities for human flourishing. Using digital technologies can create forms of dependency or a pressure to adapt, closing off other previously established options in the process.

64) This may occur insidiously and in part unconsciously as a result of behavioural shifts, without any intention on the part of the actors involved. Setting replacement as the end goal when assigning tasks that were previously accomplished by humans to technological systems, on the other hand, occurs intentionally. In and of itself, this sort of transfer already expresses one perception about human authorship. The main ethical question is whether and how it will impact the possibilities of *other* people, especially those directly impacted *by* a machine decision. This makes it necessary to render the transfer of human activity to AI systems transparent, also to those persons affected by it, and to consider *for whom* a given application entails opportunities or risks and expansions or reductions of authorship. This involves questions of social justice and power.

65) Psychological effects related to AI systems also deserve consideration, especially automation bias. People will often trust results generated by algorithms or automated decision processes more readily than those made by humans, with the effect that responsibility – at least unconsciously – is delegated to the former as "quasi-actors". Even in situations that place tight normative restrictions on an AI system and limit its role to decision-making support, automation bias can lead the system to gradually assume the role of actually making the decisions, eroding human authorship and responsibility in the process.

>> PART II: SELECTED APPLICATIONS AND SECTOR-SPECIFIC RECOMMENDATIONS

Medicine

66) The healthcare system is one field in which digital products supported by AI are used with growing frequency. Parsing the attendant opportunities and risks requires at least a threefold differentiation. First, various groups of actors must be distinguished from each other, each with different roles and responsibilities regarding the use of AI. Second, the healthcare sector encompasses various areas of application for AI products, from research to actual patient care. Finally, there are varying degrees of replacement for human activity.

67) Developing AI components that are suitable for medical practice already demands close interdisciplinary collaboration between experts and high quality standards for the training data, so as to minimise avoidable forms of bias in the results from the outset. AI systems should be designed to include plausibility checks during the use phase, to steer clear of automation bias. Adequate testing, certification and

auditing measures should be put in place to ensure that all AI systems have been sufficiently tested before they are used, and that their basic functions can be explained and interpreted by anyone who may later use them, at least in the case of systems proposing decisions that may hold serious consequences for people.

68) AI holds out a variety of advantages to the field of medical research, provided that study participants and their data remain protected. AI can play a valuable preparatory or supplementary role by searching literature and large databases or by uncovering new correlations between certain phenomena and then making accurate predictions on this basis, about how a virus might spread, for example, or the structure of a complex molecule.

69) AI instruments are also increasingly used in medical care for diagnostics and treatment, for instance for breast or prostate cancer. In this case, decision support systems model and automate decision-making processes by analysing various parameters in laboratory diagnostics, by processing images and by automatically reviewing patient records and scientific databases. Improvements in AI supported image recognition in particular have opened up new prospects for early detection, localisation and characterisation of pathological changes. The use of surgical robots is one example of AI in medical treatment.

70) Assigning a physician's activities to technology on such a small to medium scale can result in the earlier detection of tumors and a wider range of options for treatment. This increases the likelihood of successful treatment. Furthermore, the technology gives medical personnel a respite from monotonous routine tasks, and more time to interact with their patients. Such opportunities do not come without risks however. Medical professionals are liable to lose a number of skills if certain tasks continue to be delegated to technological systems, for example, or automation bias may lead them to neglect their duty of care when using AI-supported technology.

71) Taking full advantage of what AI has to offer in clinical situations and minimising the risks requires keeping sight of different levels at the same time. Among other things, comprehensive and (as far as possible) uniform technical equipment, personnel training and continual quality assurance measures should exist alongside strategies for ensuring that findings based on AI-supported protocols are also tested for plausibility, and that a patient's individual situation is given full consideration and communicated confidentially. The large amount of data that a majority of medical AI applications require likewise presents challenges, both in terms of protecting the privacy of data subjects as well as with regard to the sometimes very restrictive individual interpretation of applicable data protection regulations, which can stand in the way of realising AI's potential in clinical practice.

72) Psychotherapy is one of the few areas of medical practice in which AI-based systems have, in some cases, either largely or completely replaced physicians or other healthcare personnel – at least de facto. Tools have been in development or use in this context for years, mostly in the form of screen-based apps that offer a type of therapy supported by algorithms and are often freely available. On the one hand, the low barriers and ready accessibility of the apps may introduce people to therapy who it would otherwise reach too late or not at all. At the same time, a lack of effective safeguards for maintaining quality and people's privacy presents cause for concern, as does the possibility that people will develop emotional ties to a therapy app. There is also a controversial debate as to whether the increased use of such apps will further shrink the number of licensed therapists.

73) In light of these considerations, the German Ethics Council has formulated nine recommendations for using AI in the health sector.

   - *Medicine – recommendation 1:* Developing, testing and certifying AI products intended for medical use requires closeknit collaboration with responsible authorities and especially with professional

medical societies, both to identify weak points at an early stage and to establish high quality standards.

- *Medicine – recommendation 2:* When selecting training, validation and test data sets, measures that go beyond current legal provisions should be adopted to ensure that relevant factors for a given patient group are sufficiently accounted for (e.g. age, sex, applicable ethnic factors, pre-existing conditions and comorbidities). These include monitoring and precise yet reasonably implementable documentation requirements.

- *Medicine – recommendation 3:* When designing AI products for decision support, it must be ensured that the results are presented in a way that makes the dangers of automation bias transparent, counteracts them, and emphasises the need for a reflexive plausibility check of the course of action proposed by the AI system.

- *Medicine – recommendation 4:* When collecting, processing or sharing health-related data, strict requirements and high standards for information, data privacy and protecting individuals' personal lives must be observed. The German Ethics Council refers readers to recommendations it developed in 2017 in its Opinion on big data and health, which take their cues from the concept of data sovereignty. This notion applies equally for AI in the medical field.

- *Medicine – recommendation 5:* In cases where careful empirical study has demonstrated an AI application's superiority over conventional methods of treatment, the former must be made available to all relevant patient populations.

- *Medicine – recommendation 6:* Proven superior AI applications should be rapidly integrated into the clinical training of medical professionals, both to prepare for their expanded usage, and to

responsibly design this usage to allow as many patients as possible to benefit and to remove existing access barriers to new forms of treatment. This, in turn, will require new, pertinent curricula or training modules for basic and continued education. Other healthcare professions should similarly incorporate such elements into their training to strengthen overall user competence with AI applications in the healthcare sector.

- *Medicine – recommendation 7:* In situations where AI components find routine use, it is critical that clinical users possess a high level of methodological expertise in interpreting the results, and adhere to strict diligence requirements when collecting and/or passing on data or testing machine recommendations for plausibility. Particular attention needs to be paid to the risk that medical personnel may suffer a loss of theoretical knowledge or haptic, practical experience and any other associated skills (deskilling); this should be counteracted by effective and targeted measures for further training.

- *Medicine – recommendation 8:* As AI components take over more and more medical, therapeutic and caretaking activities, it is not enough simply to inform patients in advance about any circumstances relevant to their treatment decisions. An effort must also be made through targeted communication to preserve trust between the parties involved, and actively combat a scenario in which patients come to feel increasingly under the threat of objectification. The more human activity is replaced by technology through AI components, the greater the need to educate and support patients. The increased use of AI components in healthcare must not lead to talking medicine being further devalued or staff reductions.

- *Medicine – recommendation 9:* Fully replacing doctors with AI systems would jeopardise patients' well-being and is not justifiable

even with reference to the severe staff shortages a number of medical fields are currently experiencing. More than others, complex treatment situations require that patients have access to a personal counterpart. That counterpart may well receive increasing assistance from technology, but that does not make them superfluous as the person responsible for planning, providing and overseeing treatment.

Education

74) School education is a second field in which digital technologies and algorithmic systems are finding increased use, offering the potential for both greater standardisation and personalisation in the learning process. Potential applications range from narrowly defined, individual services to scenarios in which teaching and learning systems supported by AI intermittently or fully replace teachers.

75) The underlying concept of education in this Opinion turns on a human capacity for free and rational action that cannot be reduced to behaviourist or functionalist models. Education requires learners to develop orientational knowledge as a baseline condition for reflexive judgement and the ability to make decisions, a process that encompasses cultural learning as well as emotional and motivational aspects. Teaching and learning should both be regarded as dynamic, interactive processes that involve other people. When using AI-supported instruments in schools this makes it necessary to inspect whether a given application aligns with and promotes an understanding of people as capable of self-definition and assuming responsibility or rather hinders it.

76) The point of departure for most AI applications in education is collecting and analysing large quantities of data from learners, and occasionally teaching staff. This, in turn, raises questions about the degree

and extent to which data collection makes sense, as well as what kind of uses are desirable. In essence, data should be used to provide students with the best possible support in their individual learning, while at the same time preventing that data from being misused to track or otherwise stigmatise individual learners.

77) Collected data ideally provides a basis for individualised feedback about a person's learning or teaching, as well as appropriate responses or recommendations from the software system. By analysing learning speed, common errors or strengths and weaknesses, for example, the software can learn to recognise a particular learning profile and adjust the teaching material accordingly. While data can thus shore up the subjective impressions a teacher might have, in certain cases they might also correct them.

78) As in the medical field, using AI in school can also lead to limited, moderate or broad replacement of certain human activities and interactions. Using a software system for a precisely defined learning module constitutes one example of limited replacement. More extensive and data-intensive intelligent tutoring systems, meanwhile, are able to transmit more complex learning material in different disciplines in collaboration with learners. This allows them to cover a broader range of aspects related to teaching or in certain cases fully assume the role of a teacher.

79) Efforts are also currently underway to employ AI in analysing classroom behaviour (classroom analytics) in order to comprehensively document and understand the dynamic of entire groups of learners. The wide range of data that these methods rely on to do so, including information about student and teacher behaviour, makes them controversial. The prospect of improving pedagogical and didactic methods stands opposed to the negative consequences that widespread data collection may have for people's private lives and autonomy.

80) One particularly controversial facet of classroom analytics involves attention monitoring or affect recognition in the classroom, especially when it is based on analysing video or audio data taken from the classroom itself. While these efforts may well be linked to the goal of improving learning results, doubts persist as to whether current technology is, in fact, capable of measuring attention and emotion with sufficient precision and reliability, and without systematic bias. Moreover, the cited risks associated with gathering the necessary data are considered particularly serious in this case.

81) Overall, opportunities associated with AI in schools include personalised learning methods and alleviating teachers' workload, the possibility of greater objectivity and fairness when evaluating progress and improving access and opportunities for inclusion for learners with special needs. Beyond concerns about bias or encroaching on people's privacy and autonomy, further potential risks include growing isolation and loneliness on the part of students, as well as qualitative shifts in the nature of learning itself. The use of AI, for example, might fundamentally impact students' motivation to learn or their ability to solve more complex tasks.

82) While AI-supported teaching and learning systems are thus capable of aiding the learning process, they cannot take the place of person-to-person transmission or satisfy the more personal elements of education. Nor should the importance of school as a social space for human interaction go underestimated. Education is not simply an optimisable or calculable process of accumulating knowledge. More than anything, it consists in cultivating a constructive and responsible relationship with received knowledge. As such, special care must be taken not to diminish learning processes that are central to the development of human personality when delegating aspects of teaching and learning to machines.

83) Against this backdrop, the German Ethics Council has formulated eleven recommendations for using AI in school education.

- *Education – recommendation 1:* Digitalisation is not an end in itself. When and how AI is used should not be guided by technological visions but by fundamental notions of education, including ideas about the formation of personality. Consequently, any tools should be used in a controlled way within the educational process, and construed as one element within the relationship between teacher and student.

- *Education – recommendation 2:* Any time AI finds use in the classroom, it is essential to carefully weigh the opportunities against the risks. Protecting the autonomy and privacy of both teachers and students should rank paramount. AI presents particular opportunities in the areas of inclusion and participation, where its potential should be used to dismantle for example linguistic or physical barriers.

- *Education – recommendation 3:* Tools that replace or supplement individual elements for teaching and learning (narrow substitution) and demonstrably expand the abilities, skills or social interactions of the people using them, such as some intelligent tutoring systems or telepresence robots for remote learning situations, are in principle less problematic than those that replace more extensive or broader parts of the educational process. The greater the degree of replacement, the more rigorous the evaluation of areas of application, environmental factors and potential benefits must be.

- *Education – recommendation 4:* It is essential to establish standardised certification systems that use transparent criteria for successful learning in the comprehensive sense given above to assist school authorities, schools and teachers in deciding for or against

using a given AI product. In this context, the present Opinion can endorse the proposal in the report from Germany's Standing Scientific Commission on Education Policy (Ständige Wissenschaftliche Kommission der Kultusministerkonferenz) on digitalisation in the educational system to establish permanent digital education centres across federal states.

- *Education – recommendation 5:* Developing, testing and certifying AI products for use in school education requires closeknit collaboration between relevant authorities, professional pedagogical societies and participation of stakeholders in order to identify weak points at an early stage and establish high quality standards. Known challenges associated with AI technologies such as bias or tendencies to anthropomorphise the technology should be kept in mind during development and standardisation.

- *Education – recommendation 6:* Improving overall competence with AI technologies, especially among teachers, is an essential aspect of ensuring their responsible use in education. This, in turn, means creating and establishing apposite learning modules and curricula for basic and continued professional training courses. The risk that AI will result in a narrower pedagogical approach and a loss in teaching skills deserves particular, proactive consideration. By the same token, digital skills among learners and parents alike should be strengthened and expanded to include those that involve AI.

- *Education – recommendation 7:* AI-based tools should, as a matter of principle, also be made available to learners for self-study to promote participatory justice.

- *Education – recommendation 8:* Introducing AI tools in educational settings requires that a number of adjacent research fields are further developed, including research into theoretical foundations

and empirical evidence on its effects, for example on skill acquisition (e.g. problem solving) or on influencing child and teen development. Not only should there be greater investments in research and product development, but above all also increased practical testing and evaluation in everyday school life.

- *Education – recommendation 9:* Applying AI in educational settings further raises the issue of data sovereignty. On the one hand, this means adhering to strict requirements for protecting people's privacy when collecting, processing or sharing education-related data. Yet gathering and using big data responsibly and in a way geared toward the public good should also be an option, as is the case with prognostic methods that support teaching.

- *Education – recommendation 10:* Completely replacing teachers with AI systems goes against the concept of education presented in this Opinion, nor can it be justified by pointing out current staff shortages or poor teacher training in certain fields. In the complex situation of school education, there is a need for a personal counterpart. That counterpart may well receive increasing assistance from technology, but this does not make them superfluous as the person responsible for providing pedagogical support or evaluating the learning process.

- *Education – recommendation 11:* Overall, the members of the German Ethics Council are sceptical about using audio and video monitoring in the classroom, both for the epistemological and ethical challenges it presents and when weighing the possible benefits and harms. Especially the analysis of attention and emotions in the classroom via audio and video monitoring with currently available technologies does not seem justifiable. Some members of the Ethics Council do remain open to monitoring attention and emotion in the future, provided that the gathered data can be shown to improve the learning process in a scientifically

demonstrable way, and that any monitoring necessary to gather data does not have an unacceptable impact on the privacy or autonomy of the learners and teachers under observation. Other members of the Council, by contrast, consider the implications for privacy, autonomy and justice unacceptable in general, and advocate banning attention monitoring technologies in schools.

Public communication and opinion formation

84) Transformations in the digital world are also affecting the realm of political communication. The rapid spread of digital platforms and social media, with the array of information and communication patterns that their algorithms present, affects not only individual social spheres but potentially also large parts of public communication and opinion formation. This, in turn, holds consequences for democratic legitimisation structures.

85) Many of these platforms now offer a similar range of options for creating and distributing multimedia content, interacting with other users or reacting to the content they post, and searching for or subscribing to material. Users also have a wide range of options to create targeted advertising for their content, offer products or services directly or make purchases. Nearly every platform or service of this sort is operated by private companies from the USA or China, with the largest social networks run by a small handful of businesses. Today, the companies' market power, paired with the versatility and broad integration of their services, mean the platforms have come to serve as vast socio-technological infrastructures upon which a majority of online behaviour plays out – all according to the provisions of a small number of businesses.

86) The wealth of information and interactive opportunities social media has to offer come accompanied by technical challenges as well as

economic possibilities, which have jointly contributed to how platforms currently function, and the business models on which they are based. The amount of content, meanwhile, confronts platforms and users alike with the issue of choosing information. At present, this task is delegated almost exclusively to algorithms, which subsequently ensure that visitors to a platform are shown a specific sequence of content that has been individually tailored to them.

87) The criteria by which algorithms decide which content will be shown are tightly bound up in economic factors. A majority of platforms and services follow an advertising-based business model. It works best by pinpointing users' individual interests and having them spend as much time on the platform as possible, whilst presenting them with advertising attuned to their personal interests. This gives platforms an interest in gathering as much data as possible about users' personal backgrounds, interests, usage patterns and social network, and then using the data when selecting personalised content (profiling).

88) Allowing algorithms control over the individualised information users are shown, a process that is tightly wrapped up in economic and attention-based considerations and is continuously being adapted based on usage patterns, leads to content that is especially sensational or which triggers intense emotional reactions spreading with disproportionate speed and breadth. Among other things, it facilitates the spread of false news or content that includes hate speech, libel or incitement to hatred.

89) Platforms have responded to the challenges of potentially problematic and viral content by attempting to moderate it according to different criteria (content moderation), an effort that relies on humans as well as algorithmic systems. These criteria derive their basis from legal frameworks as well as platform-specific rules of communication, which may even lead to the deleting, blocking or containing the spread of content that is actually legally permissible.

90) Human moderation is typically accomplished by employees working for third parties that have signed a contract with a given platform. Often working under precarious conditions, these employees are regularly exposed to extremely distressing material such as executions, child abuse, animal torture and suicide. What is more, they often have only a few seconds to take in linguistically and culturally complex nuances that can play a decisive role in a given post's acceptability.

91) Algorithms, by contrast, are able to filter out offensive material without humans having to view it, and are better able to handle the incalculable amount of data and content available on the Internet. However, the automated methods that exist to date are often still incapable of fully appreciating a post's cultural and social context, and judging it fairly. The current structure of legal incentives gives rise to the risk that even content that does not violate rules will be systematically deleted or otherwise made inaccessible (overblocking).

92) Human capacities for action can be expanded or diminished in different ways through the described functionalities of platforms and the socio-technological interconnections that unfold. On the one hand, assigning algorithms to curate or moderate material can make the process easier and more efficient, expanding peoples' options for action by enabling them to access information or attain personal goals more effectively or quickly, for example, or granting them greater scope for other activities by delegating content selection to algorithms.

93) The capacity for action and people's personal freedom may be diminished on the other hand in cases where it proves difficult to resist being drawn in by online material or set a healthy limit to its use. Allowing algorithms to curate content can also diminish human authorship by anticipating certain decisions regarding relevance, limiting the extent to which we are able to reason out alternatives for ourselves in the process.

94) Beside their more general effects, the ways in which platforms and social media operate also impact two aspects critical to the process of forming public opinion – the quality of information and the quality of discourse – with potentially far-reaching consequences for the political process. No final conclusions can be drawn at present as to just how widespread and potent the effects discussed below are, in part because the data are at times unclear or contradictory. It is nevertheless worth taking a closer look at the mechanisms described above, if only because the processes they touch on are foundational to our democracy.

95) Where quality of information is concerned, it is at first worth pointing out a positive development in the growing number of available sources. Frequently pitted against this are worries about the negative results that current practices of algorithmic curation can have, whether it is promoting the spread of false news or conspiracy theories, contributing to the creation of filter bubbles and echo chambers or prioritising content that provokes harmful emotional and moral reactions and interactions.

96) Uncertainties do persist about the actual extent of these phenomena. Yet it seems plausible that false news, filter bubbles and echo chambers, as well as the heightened emotional and moral tone of a great deal of material available online, could adversely effect the quality of information. Under these circumstances, the sheer power of the algorithms in use imposes a de facto limit on our freedom to search out high-quality information.

97) The quality of discourse is also impacted, both from an ethical and political perspective, by changes in the quality, presentation and dissemination of information as mediated by algorithms. Here too, positive developments and possibilities present themselves, especially in the much wider range of opportunities for participation and direct networking that platforms and social media offer. Yet opportunities

for the quality of discourse also stand opposed to more negative developments. The focus is on three main topics: the political polarisation of public discourse, political advertising and manipulation and a field of tension defined by increasingly coarse discourse on the one hand, and excessive interference in the freedom of expression and opinion on the other.

98) There is a great deal of evidence that the relative ease with which emotionally and morally charged content spreads, has in part shifted the tone of discourse, including and especially via channels that are actively involved in shaping political discourse. When, for example, Facebook altered its criteria for selecting material to give wider distribution to content that generated an especially vehement reaction, many political communications teams altered the tone of their posts to meet the new standard.

99) Online platforms also present a tremendous opportunity for particularly powerful communications campaigns, which can operate inconspicuously in the day-to-day digital environment without users being particularly aware of them. The wealth of data stored in the profiles that result from people's usage patterns can further be used to place political advertising with pinpoint accuracy (targeted advertisement), strategically disinform people or dissuade them from voting. While the actual success rate of this sort of microtargeting has not yet been sufficiently researched, the mere knowledge that attempts are being made to manipulate political preferences on the basis of highly personal psychological characteristics is liable to harm political discourse, and undermine trust in political processes of opinion formation.

100) Fake accounts used to wield strategic influence over political discourse, some of which operate automatically (bots), can further undermine trust. Communication campaigns can use such fake profiles

to successfully augment their messages, lending them a greater power to convince and thereby distort discourse in problematic ways.

101) The aforementioned trend towards an increasingly vitriolic tone on platforms and in social media is accompanied by the concern that a rise in highly negative and aggressive styles of communication, including hate speech, threats and calls for violence can contribute to a brutalisation of political discourse. Even if smear campaigns spread online do not wind up as actions taken in the real world, they can have a chilling effect on discourse, for example in cases where the opinions voiced cause so much discomfort and fear that others refrain from participating in public discourse, thereby diminishing their freedom and options for action.

102) On the other hand, efforts to mitigate potentially problematic content through moderation escape criticism, as they raise their own questions of democratic theory. Excessively deleting or blocking content can constitute interference in freedom of opinion and freedom of the press, and lead to chilling effects of its own in the event that people refrain from publishing content in the first place, either out of fear that it may be immediately deleted or lead to their accounts being (temporarily) shut down.

103) All in all, the phenomena and developments that are coming to pass within the socio-technological environments created by digital networks can strongly impact processes of public communication, as well as political opinion formation and decision-making, including – and perhaps especially – in democratic societies. Against this backdrop, the German Ethics Council has arrived at ten recommendations for this field of AI application.

   ›› *Communication – recommendation 1:* Regulating social media: There should be clear legal guidelines for the form and extent to which social media and online platforms are required to provide

information about how they curate and moderate content, and explain how this is implemented based on institutional regulations. This process must be subject to external verification. Purely voluntary efforts by private actors, especially non-binding reviews by self-appointed supervisory bodies, are not sufficient. Regulatory approaches for this already exist in the European Union's Digital Services Act, but they do not yet go far enough.

- *Communication – recommendation 2:* Transparency regarding moderation and curation practices: Instead of general guidelines on moderation or deletion and uninformative figures about the number of deletions, external bodies must be able to follow the means, circumstances and criteria by which these types of decisions are made and carried out, and the role that algorithms or human moderators have assumed in the process. Moreover, the basic mechanisms for determining how content is curated on social media and online platforms must be disclosed to an extent that allows systematic bias and potentially resulting informational dysfunctions to be identified. The reporting requirements and transparency guidelines prescribed by Germany's State Media Treaty (Medienstaatsvertrag), Network Enforcement Act (Netzwerkdurchsetzungsgesetz) and the Digital Services Act are not currently sufficient to ensure this. A number of the reporting requirements laid out under Articles 12 ff. of the General Data Protection Regulation apply only at the national level, and often do not cover these more extensive aspects.

- *Communication – recommendation 3:* Research access to platform data: To investigate the impact of platforms and social media, their influence on public discourse, but also other topics of high social relevance, it should be ensured that independent researchers are not denied access to relevant platform data by blanket references to trade or business secrets. Secure methods of access that conform to data privacy laws and uphold research ethics will need to

be found. The Network Enforcement Act and the Digital Services Act already contain regulations for accessing data, though they are too limited in their scope. The Data Act provides a comparable regulatory framework.

- *Communication – recommendation 4:* Addressing security, data privacy and confidentiality concerns: Disclosure or data access requirements must be tailored to the context, whilst adequately addressing requirements for security and protection against misuse, violations of data privacy, intellectual property and trade secrets. Depending on the context, a distinction must be drawn between more or less clearly defined auditing times and levels of disclosure.

- *Communication – recommendation 5:* Personalised advertising, profiling and microtargeting: Personalised advertising constitutes the central business model of social media and online platforms. Profiling and microtargeting can have negative consequences in the context of public communication and opinion formation however, especially in the case of political advertising. In order to prevent such negative effects through effective regulation, it is first necessary to create conditions that will enable exploration and inspection of the relationship between business models and practices of algorithmic curation with respect to their modes of operation and effects. The current proposal for an EU regulation governing transparency and targeted political advertisements addresses this need. By the same token, this also reveals the challenges of shaping laws that are simultaneously effective and do not impinge on free political discourse.

- *Communication – recommendation 6:* Improving the regulation of online marketing and data trading: Many dysfunctional patterns in information and communication find their origin in online marketing which, in turn, constitutes the primary business model for many forms of social media and online platforms. Online

marketing operates by collecting, analysing and then selling a wide range of data about the people using the services. The problem is not funding from advertising per se so much as the invasive way in which data are handled. On the one hand, this makes it necessary to research the business model's impact on public discourse more closely. On the other, there is a need for improved legal regulations that will more effectively protect the basic rights of individuals while minimising any negative systematic effects on public discourse. In 2017, the German Ethics Council already made suggestions in this direction under the concept of "data sovereignty" in its Opinion on big data and health. While European regulations like the Digital Markets Act address the problem of the data power of large platforms, they do not do so with a view to the implications for public discourse, if only for reasons of regulatory competence.

- *Communication – recommendation 7:* Power constraints and control: Companies with a de facto monopoly of power in how data or facts are presented publicly must be made to protect pluralism and minorities and guard against discrimination through legal requirements and appropriate monitoring. Some members of the German Ethics Council believe that existing media law regulations for ensuring plurality, neutrality and objectivity should be expanded to apply generally to news services on social media and online platforms, to the extent that they resemble those of traditional media.

- *Communication – recommendation 8:* Expanding user autonomy: Platforms and social media should also make content available without personalised curation. Additionally, users should have a wider range of options at their disposal for choosing the criteria by which algorithms select content on platforms and social media, and the order in which it appears. This should include an option allowing users to view opposing positions, which go against the

preferences they have articulated to date. Different options for viewing content should be easily visible and accessible.

- *Communication – recommendation 9:* Fostering a critical relationship towards content: In order to curb the thoughtless spread of questionable content, platforms should develop and employ a broad system of notifications that encourage users to engage critically with material before deciding to share it or respond publicly. They might be queries as to whether the user has read a text or watched a video before sharing it or information about the legitimacy of a given source.

- *Communication – recommendation 10:* Alternative information and communication infrastructure: Consideration should be given to establishing a public European digital communications infrastructure alongside existing private social media services, the operation of which would not be geared towards commercial interests such as having people spend as much time on the platform as possible. The aim here would not be to extend public service broadcasting (TV and radio) to another digital platform, but to provide a digital infrastructure as an alternative to commercially-driven, oligopolistic services. An organisation in the form of a public foundation could be created to assure the new platform retained sufficient independence from the state.

Public administration

104) For many individuals and organisations, the field of public administration – whether encountered through the public finance sector, the tax system, registration requirements, social services or in offender and juvenile court assistance – constitutes a direct experience of the power of the state. A functioning, transparent administration that is recognised as legitimate and responsive to its citizens is essential

for a well-functioning community and the acceptance of democracy and the state. Digitalisation strategies in this field combine hopes of rationalising and accelerating the pace of state administrative action and developing more effective and coherent strategies for using data, at the same time expanding opportunities for researchers and citizens to share their expertise. Opposed to this aspiration is the dystopian vision of an "algocracy", in which autonomous software systems exercise state rule over humans.

105) Automated decision-making systems (automated or algorithmic decision-making systems, ADM systems) are increasingly finding widespread use in public administration. Examples include evaluating a person's chances on the job market, reviewing or allocating social services or making predictions in police work. One area of particular interest in this context is the extent to which AI systems influence humans' ability to act and human authorship. The frequency with which we tend to accept machine recommendations unreservedly (automation bias) already grants software used in assisting with administrative decisions a far-reaching impact.

106) Using AI in public administration raises other questions as well, especially pertaining to justice. Examples include whether and to what extent the systems used actually improve diagnoses and prognoses, whether the same accuracy rates exist for different fields of application or groups of people, and the possibility of systematic bias or discrimination (algorithmic bias). At the same time, data-based systems are also capable of uncovering historical injustices or human prejudice, allowing for remediation.

107) Normative conflicts between objectives or rules that are irreconcilable under German law – which is based on deontology – constitute a fundamental boundary in the use of automated decision-making systems. In the German legal system, it is never weighing the consequences alone that determines what is lawful. It is the unconditional

rights to the protection of the person that must be upheld, resulting in limits placed on delegating ethical and legal decision-making to algorithms.

108) Social services constitute one administrative area in which decisions with far-reaching consequences for those affected are made, for example about whether to grant state assistance, take action where a risk to child welfare may be present or assess the risks posed by offenders in probation services. The algorithms increasingly used to support these decisions are, in fact, capable of expanding professionals' competence, helping specialists set otherwise intuitive assessments on a more solid footing with data or correct them when necessary. The result: standardised decisions based on evidence. This is particularly important when assessing the potential for risk, for instance when a threat to child welfare is suspected or in a probation service.

109) Yet AI assistance in providing pertinent results can also diminish human authorship. This might happen in a professional context in the event that an algorithm's suggestion is taken up without first checking it (automation bias). People affected by those decisions may also be negatively impacted if algorithmic recommendations are tinged with bias, by being unjustifiably deprived of opportunities for action or development.

110) Especially when determining a person's need for assistance, using algorithmic systems conceals the risks of abandoning a dialogic relationship, which can be critical for the person to experience self-efficacy. If the personal dimension in determining an individual's level of need is neglected as a result of the algorithm-based computerisation of social services, the positive effects even of material assistance can quickly fizzle out, with scarcely any lasting effect. Austria's AMAS system, for example, predicts a person's prospects for rejoining the job market. It has been criticised for prioritising the values, norms and goals of a restrictive fiscal policy that runs diametrically opposed

to the goals of a welfare system based on people, which focuses on each individual's need for assistance.

111) Crime prevention is another area where algorithms are used more and more frequently for risk analyses. Predictive policing relies on corresponding applications to assist with preventive police work by forecasting future crimes, the people who are liable to commit them, and places where they may be committed in order to prevent crimes. The debate around procedures related to individual people is especially controversial in this context. On the one hand, there is the hope for better police work and the more effective protection of potential victims. On the other hand, algorithmic error or bias in crime prevention efforts can have especially grave consequences for people who have been unjustifiably classified. Moreover, it can have a particularly broad impact due to their systemic embedding in the software.

112) Ensuring data privacy poses a further problem for predictive policing. As a general rule, the data used in police work are particularly sensitive. Particularly with regard to what are known as chat controls to prevent and combat the sexual abuse of children, for which the European Commission proposed a regulation in May 2022, questions are being asked as to whether the groundless and blanket surveillance of private communication is justifiable or instead constitutes a disproportionate encroachment on fundamental rights.

113) Concerns include the eventuality that allowing algorithms to guide police work runs the risk of fixing a mechanistic image of people that objectifies the individual and reduces them to data-driven classifications, all the while paying scant attention to the overall societal causes of criminality.

114) Overall, the use of automated decision-making procedures in public administration presents new opportunities and challenges. These, in turn, prompt deeper ethical questions as well as questions of

democratic theory, for example in terms of how comprehensible, explainable and trustworthy administrative action is, but also in the shape of concerns about discrimination and technocracy, where human communication and deliberation disappear behind anonymous reams of data and standardised user interfaces.

115) To the extent that algorithmic systems grant access to large quantities of data whose targeted evaluation provides a firmer footing for decisions, such systems can support human authorship and are, in principle, justified from an ethical standpoint. At the same time, uncritical acceptance of a system's recommendations threatens to diminish human authorship, including a worst case scenario in which all that remains is an automated process by means of which technical systems arrive at far-reaching, potentially existential conclusions for those concerned, and where systemic errors or bias can no longer be identified.

116) When using AI in public administration this makes it essential to carefully assess and weigh in detail and in relation to the context the effects that a given measure will have on the authorship of all those involved or otherwise affected, to see what conflicts may emerge and how they might or ought to be dealt with. This has led the German Ethics Council to nine recommendations.

- *Administration – recommendation 1:* The more forcefully a given decision will impact an individual's legal position, the greater the need to inspect the increased standardisation and blanket categorisations of individual cases associated with automated decision-making (ADM systems), and supplement them with considerations specific to the case at hand.

- *Administration – recommendation 2:* To preempt the evident danger of automation bias, precautionary technical and organisational instruments should be established that make it difficult

for professionals to accept algorithmic recommendations unseen, even for those with the final say. It must be examined whether reversing the obligation to give reasons (not a deviation, but compliance is to be justified) might serve as an appropriate precautionary measure.

- *Administration – recommendation 3:* State institutions' commitment to upholding constitutional rights means a high bar must be set for transparency and comprehensibility requirements when using and developing algorithmic systems, both to protect against discrimination and meet the institutions' legal obligations to justify their decisions.

- *Administration – recommendation 4:* Quality criteria for software systems (e.g. with regard to accuracy, error avoidance, absence of bias) used in public administration must be established in a binding and transparent fashion. It is also necessary to document any methods used. In this regard, procurement practices by which state authorities currently purchase software systems should also be subject to critical evaluation.

- *Administration – recommendation 5:* Wherever algorithmic systems are used in public administration, it is to be ensured that the people using the systems possess the necessary skills. Beyond knowing simply how to use the systems, this includes a familiarity with their limitations and potential sources of bias, so that the systems can be implemented responsibly.

- *Administration – recommendation 6:* Affected persons' rights to inspect and object to decisions must also be effectively guaranteed when algorithmic systems are used. This may require further effective procedures and institutions.

- *Administration – recommendation 7:* In the public, political and administrative spheres, awareness should be raised of the potential dangers posed by automation systems, such as violations of privacy or forms of systematic discrimination. This includes a public debate about whether a technical solution is even necessary in certain contexts.

- *Administration – recommendation 8:* In the area of social services, it must be ensured that ADM systems neither undercut nor supplant basic standards of social-professional interaction (e.g. joint social diagnoses or planning assistance as *part* of treatment and support measures). That especially includes measures to prevent an oversimplification of individual case constellations and prognoses through ADM-induced coarse categorisation of case and/or beneficiary groups. Care must be taken here that determining an individual's need for assistance does not become any more complicated, and that the process of identifying those needs, as social legislation requires, does not gradually give way to one-sided external interests in minimising risk or keeping costs down.

- *Administration – recommendation 9:* Emergency management authorities, among them the police, work in areas with a particularly delicate relationship to basic rights. This, in turn, affects the degree to which it is permissible to use algorithmic systems in predictive police work. Risks such as encroaching on a person's privacy or potentially inadmissible discrimination of the persons affected by the use of ADM systems must be carefully weighed against the prospects of dramatically improving state emergency response systems, and brought into a balanced relationship with one another. Any social deliberation necessary to do so should be conducted at a broad level. The debate must take account of the difficulty in determining the relationship between freedom and security. Preventing any violation of the law would not be possible by means of the rule of law.

>> PART III: OVERARCHING TOPICS AND RECOMMENDATIONS

Preceding analysis in summary

117) The concept of artificial intelligence has drawn increasing attention within public debate. At times it is linked to excessive hopes, at others to excessive fears. The German Ethics Council draws a fundamental, normative distinction between humans and machines. Software systems possess neither theoretical nor practical reason. They do not act or decide for themselves, nor can they assume responsibility. Even in cases where they may simulate empathy, a willingness to cooperate or the capacity for insight, they are not personal counterparts.

118) Human reason is always bound up within a concrete shared social environment; that is the only way to account for it becoming effective. The individual acts rationally as part of a shared social and cultural environment. This is reason enough why neither theoretical nor practical reason can be ascribed to the software systems addressed in this Opinion.

119) While people develop digital technologies and use them as a means to pursue human ends, these technologies subsequently impact human options for action. This may result in new opportunities, but can also necessitate unwanted adaptations. Machines thus exert a strong influence on humans' capacity to act, even if they are not acting themselves, and can considerably expand or diminish human options for action.

120) Ultimately, delegating human activity to machines should aim to expand human capacity for action and authorship. Diminishing, dispersing or evading responsibility, on the other hand, must be avoided. To that end, any transfer of human activities to AI systems should be undertaken with sufficient transparency for all those involved to be able to understand the key elements, parameters and conditions of a given decision.

121) Arriving at an ethical assessment of the value and benefit of delegating previously human action to machines will always entail considering the situation at hand, taking equal account of the various perspectives of the different parties and the long-term impact of such transfers. As is so often the case, the challenges lie in the details, more precisely in the details of the technology and its intended context, and of the institutional and socio-technological parameters.

122) To facilitate this context-specific view, the German Ethics Council has in this Opinion looked at examples of the use of AI in the fields of medicine, school education, public communication and administration. The Council consciously chose sectors that differ greatly in terms of the scale of AI penetration, each showcasing the varying degrees to which AI may replace previously human activities. All four featured sectors are characterised by highly asymmetrical relationships and power dynamics, making it all the more important for AI to be used in a responsible manner that keeps sight of the interests and well-being of particularly vulnerable groups. Taking account of

the differences between fields in how each uses AI and the degree to which activities have been delegated to machines allows for nuanced ethical consideration.

Development of overarching topics and recommendations

123) The socio-technological developments and ethical dimensions described within the four fields reveal a series of overarching topics and challenges that are relevant to all four, albeit at times in different forms. To ensure that society deals well with AI in the future with regard to an expansion of human capacity to act and authorship, these sorts of overarching questions must not only be addressed in individual areas but also in interlinked cross-sectoral approaches.

124) This type of thinking, in equal parts horizontal and vertical, presents special challenges for policy-making and future regulatory efforts. The overarching topics presented below, each of which culminates in a recommendation, should therefore serve as impetus for a wider debate about the ways in which future policy decisions and technologies can and must always take account of such broader questions simultaneously and alongside sector-specific aspects.

125) The *first overarching topic* returns to a central concept in this Opinion, namely the expansion and diminishment of human options for action. One cross-sectoral commonality with regard to the desired expansion of human options for action is that a complete replacement of human actors by AI systems is prohibited wherever concrete interpersonal encounters are a necessary prerequisite for achieving the objectives of actions. However, there is also a need to carefully consider differences when using AI within individual fields of action.

   ›› *Recommendation overarching topic 1:* The advantages and disadvantages of AI will vary considerably for different groups of

people, as will the risk that individual users run of losing certain skill sets. Therefore, planning the use of AI for different fields of action requires a differentiated approach that clearly articulates the aims and responsibilities in the given case, as well as a timely evaluation of the real-life consequences of such use to better adapt the systems to the specific context and to continuously improve them.

126) The *second overarching topic* concerns knowledge generation by AI and the handling of AI-supported predictions. A central premise in this regard is that correlations and data patterns should not be equated with explanations and justifications of the causes of events, but must also be evaluated qualitatively and assessed normatively. With probabilistic methods, residual uncertainties always remain, the acceptability of which must be decided on. From an ethical point of view, it is positive that using AI has led to considerable functional improvements within each of the four fields considered in this Opinion, with more expected in the future. Yet a fundamental and normatively problematic line has been crossed when functional improvements, in a process that may even go unnoticed, shift to replacing moral competence and the associated responsibility.

›› *Recommendation overarching topic 2:* To avoid diffusion of responsibility, using AI-supported digital technologies should be designed to assist rather than supplant human decision-making. The use of AI must not occur at the expense of effective options to control it. Especially in areas that feature greater levels of intervention, people who are impacted by algorithmically supported decisions must be granted access to the basis for those decisions. This, in turn, requires that at the end of any technical process, there is a person equipped with the power of decision who is both capable of and obligated to assume responsibility for that decision.

127) The *third overarching topic* is the risk that statistical stratification poses to the individual. Many AI applications are based on correlations that are discovered when analysing large amounts of data and can be used to assign individuals to cohorts with certain combinations of characteristics. Generating this sort of cohort and the predictions they allow algorithms to make on their basis can improve the overall quality and effectiveness of an application. Yet this can also pose problems for individuals who are affected by such collective conclusions, especially in cases where the resulting diagnosis or prognosis does not apply to them.

>> *Recommendation overarching topic 3:* Aside from any specific and immediate problems arising out of data-based software, how to protect peoples' privacy, for example, or prevent discrimination, it is essential to carefully consider the long-term effects of assigning individuals to preconfigured statistical categories, as well as any repercussions on their subsequent options for action – in terms of expanding or diminishing these options – both at the individual and collective levels, and across different sectors. Case-by-case decisions also remain an important option. Evaluations or predictions made with AI can serve as an aid when the conditions are right, but are not suitable as a means of reaching a *definitive* assessment of a situation or decision. Pragmatic and heuristic factors such as testing results for consistency against different sources of evidence or estimates of success have an important role to play in this regard.

128) The *fourth overarching topic* concerns what AI implies for human competencies and skills. Their acquisition and maintenance can be jeopardised by the delegation of human activities to machines. As with other technologies, using AI applications can lead humans to neglect or entirely lose certain capacities, creating dependence on these technologies. It presents a grave risk if those losses of human

skill sets and proficiencies occur in fields that are particularly important, even critical to society.

- *Recommendation overarching topic 4:* It must be carefully monitored whether and to what extent using AI results in unwanted losses in human skill sets. Such losses should either be kept to a minimum or compensated for by creating a sensible plan for human–technology interactions when developing and applying new technologies, by drawing on any appropriate institutional or organisational frameworks and by taking targeted countermeasures like specific training programmes. Skill sets can be lost either at the individual or collective level, making it essential to ensure that societies as a whole do not make themselves overly susceptible to (temporary) technological failure in delegating tasks to technology. Alongside these systematic aspects, any negative effects on individual autonomy or self-perception must also be mitigated.

129) The *fifth overarching topic* is the balancing act between guarding privacy and autonomy on the one hand, and the risks posed by surveillance and chilling effects on the other. The large quantities of personal data on which AI applications rely, and the possibility of using them to arrive at sensitive prognoses, do not simply represent encroachments on the privacy of the people whose data are being collected. It also leaves them vulnerable to possible discrimination or manipulation that may arise from the data being processed. In this context, the term chilling effect describes how people's concern that they are being observed, recorded or analysed impacts their behaviour.

- *Recommendation overarching topic 5:* The emergence, nature and development of the described phenomena should be subject to thorough empirical research. Suitable and effective legal and technical preventive measures must be found (privacy by design is one example) to address the issue of surveillance and any parallel risks arising from chilling effects, and to block online behaviour

and personal data from being excessively tracked and traded in. The interests of the data subjects must remain front and centre. Vulnerable groups should receive special attention, since many contexts in which AI is used are characterised by asymmetrical power relations. Care must be taken to ensure that expanding the options for action for some people does not come at the cost of diminishing those of others, especially disadvantaged groups.

130) The *sixth overarching topic* turns on concepts of data sovereignty and using data for the common good. The German Ethics Council has already elaborated on this subject in a 2017 Opinion on big data and health. The goal in this context is to find ways for AI to make meaningful use of data for a variety of important undertakings without, at the same time, improperly encroaching on data providers' right to privacy. It is an open question whether current data privacy laws and practices serve both of these aims faithfully. While justified concerns about undetected and widespread intrusions on privacy rights and informational self-determination exist for any number of fields, in other contexts interpreting data protection laws too strictly may make it either impossible or extremely difficult to achieve important social aims, regarding patient care or acquiring scientific knowledge for example, but also municipal services of general interest.

›› *Recommendation overarching topic 6:* When using AI in a given field, new ways must be found for facilitating or allowing (secondary) data use for the common good while also taking account of the challenges or benefits specific to that field or context, thus expanding options for action. At the same time, it is essential to shift consciousness among the general public as well as those directly involved in shifting data usage away from a predominately individualistic, overly short-sighted perspective, towards a position that incorporates and balances systematic notions or those of the common good. The same attitude should prevail in future policy-making and regulatory efforts, to a much greater extent

than has been the case so far. This is the only way to ensure that, alongside the risks that undoubtedly arise from the widespread use of AI, we do not lose sight of the important opportunities of responsible AI use.

131) The *seventh overarching topic* concerns critical infrastructure, dependency and resilience. In the course of digitalisation, infrastructures, such as electrical grids, are increasingly monitored and controlled via the Internet. At the same time, digital technologies themselves have become a form of infrastructure. Humans base their behaviour on the assumption of infrastructure being readily available and functional. In the course of this social appropriation, dependencies arise that can jeopardise human autonomy. The lack of transparency and comprehensibility of AI-supported systems constitutes further cause for concern as they are increasingly used to manage infrastructure; so too does the possibility of society and its institutions growing more and more vulnerable as the infrastructure systems and their operation become more complex.

›› *Recommendation overarching topic 7:* The resilience of socio-technical infrastructures must be strengthened and any individual or systemic dependencies minimalised in order to expand human actors' authorship and freedom to act. This begins by acknowledging the important role that digital technologies have to play in infrastructure, then working more diligently to protect and ensure the resilience of critical digital infrastructures, including through political action. No matter which area or field, it is essential to avoid one-sided forms of dependency that leave society vulnerable and open to attack in moments of crisis. Reducing dependency on digital technology requires users having the option to choose between alternatives without any major losses in functionality. This includes the need for interoperability, being able to switch easily between different systems. Setting this goal makes it particularly important to establish or expand alternative infrastructures. In the

context of public opinion formation, there is a pressing need to establish independent and public digital communication platforms. In other sectors such as administration, education and medicine, relying too heavily on a handful of systems or actors also potentially reduces the individual and collective ability to act.

132) The *eighth overarching topic* revolves around path dependencies, secondary usage and the dangers of misuse. Path dependencies arise when decisions made at the beginning of a certain line of development continue to make their effects felt long after, at times becoming difficult to reverse even after the context may have changed. Once technologies have been introduced, there is probably also a tendency to fully exploit their potential, even beyond the original field of application. Secondary uses like this are not problematic in theory, although once established it can be difficult to prevent a technology from being used in other scenarios, including scenarios of misuse. More often than not, digital technologies and, in particular, enabling technologies such as machine learning open up a world of usage possibilities in which it becomes increasingly difficult to differentiate between what is use and what is misuse.

›› *Recommendation overarching topic 8:* In the case of technologies with a broad impact or distribution, and particularly in cases where it is either impossible or nigh on impossible to avoid using them, potential long-term developments like path dependencies in general, and dual use potential specifically, should be explicitly anticipated and incorporated into development planning as a matter of course. This applies in particular to application planning. In addition to effects that cause direct harm in a specific field, effects that transcend any one individual area should also be taken into account, even if they are of course much less tangible and predictable. High standards for security and protecting individuals' privacy (for instance security by design and privacy by design) can likewise help to limit, and, where possible, prevent

subsequent misuse. Especially high standards must be maintained in cases where the technology is particularly invasive, for instance public administration decisions, where citizens may have no choice but to use the technology. Open source methods may represent an appropriate way of ensuring and verifying this.

133) The *ninth overarching topic* concerns bias and discrimination. Data-based AI systems learn on the basis of pre-existing data. The resulting prognoses and recommendations continually bring the past into the future, reproducing or even strengthening stereotypes and pre-existing social inequity and injustice by incorporating them into seemingly neutral technologies. AI systems are rarely developed with an explicit intention to discriminate. Rather, discriminating effects arise out of the interaction between social realities or stereotypes and technical and methodological decisions. It is nonetheless at least conceivable that complex systems could harbour an explicit intention to discriminate.

›› *Recommendation overarching topic 9:* In light of the challenges, guarding against discrimination requires an *appropriate degree of supervision and control* over AI systems. Especially in sensitive areas, this demands the establishment or expansion of well-equipped institutions. The greater the level of intervention and the fewer options that exist for working around the system, the stricter the requirements for minimising discrimination must be. Even with technology still under development, it is essential to minimise discrimination and/or create the conditions for fairness, transparency and comprehensibility. This should be encouraged both by incentives – for instance research funding opportunities – and corresponding legal requirements, for example to disclose which anti-discrimination measures were adopted in developing a given software. Yet technical and regulatory measures for minimising discrimination have their limits, among other reasons because different fairness goals are not always technically feasible at

the same time. Ethical and political decisions must, therefore, be made as to which criteria for justice should be applied in which context. These decisions must not be left to the people developing the software or others who are directly involved. Instead, suitable procedures and institutions need to be developed in order to discuss these criteria in a context-specific and democratic manner and to renegotiate them as often as necessary. Depending on the intended use and sensitivity of the system, public engagement may be necessary. Special care must be taken to protect the most vulnerable populations or those particularly affected by decisions.

134) The *tenth overarching topic* concerns questions relates to transparency and comprehensibility, as well as control and responsibility. The frequent opacity of AI systems is linked to different factors, ranging from a wish to protect intellectual property and complex and inscrutable procedures to a lack of transparency in the decision-making structures in which algorithmic systems themselves are embedded. While transparency and comprehensibility are related to the relative degree of control and accountability that exist for using a given algorithmic system, neither aspect is mandatory or sufficient.

>> *Recommendation overarching topic 10:* There is a need to develop well-balanced, task-, audience- and context-specific standards for transparency, explainability and comprehensibility that establish their importance for supervision and responsibility and ensure their implementation by means of binding technical and organisational guidelines. In this context, the process must adequately satisfy any requirements pertaining to security or guarding against misuse or violations of data privacy, and protect intellectual property and trade secrets. Dependent on the context, different points in time (ex ante, ex post, real-time) as well as different procedures and degrees of disclosure must be specified.

135) This Opinion has considered the impact of humans increasingly delegating activities to digital technologies, especially software systems based on AI. Numerous examples from the fields of medicine, school education, public communication and opinion formation, and public administration have shown how this process of delegation is connected both to an expansion and diminishment in humans' options for action, making it either beneficial or harmful toward realising human authorship.

136) Fostering human authorship must remain the aim and guiding principle for ethical evaluation. In doing so, it is essential to keep in mind that expanding one group's options for action may involve diminishing those of another. These divergent effects must be taken into consideration especially where protecting and improving the lot of vulnerable or disadvantaged groups is concerned. While normative requirements shaping the use of these technologies – for instance requirements for transparency or comprehensibility, protecting privacy or preventing discrimination – are ultimately of great importance no matter what field or who is impacted, they must still be specified by sector, context and audience to ensure they are both appropriate and effective.

Members of the German Ethics Council

Prof. Dr. med. Alena Buyx (Chair)
Prof. Dr. iur. Dr. h. c. Volker Lipp (Vice-Chair)
Prof. Dr. phil. Dr. h. c. Julian Nida-Rümelin (Vice-Chair)
Prof. Dr. rer. nat. Susanne Schreiber (Vice-Chair)

Prof. Dr. iur. Steffen Augsberg
Regional bishop Dr. phil. Petra Bahr
Prof. Dr. theol. Franz-Josef Bormann
Prof. Dr. rer. nat. Hans-Ulrich Demuth
Prof. Dr. iur. Helmut Frister
Prof. Dr. theol. Elisabeth Gräb-Schmidt
Prof. Dr. rer. nat. Dr. phil. Sigrid Graumann
Prof. Dr. rer. nat. Armin Grunwald
Prof. Dr. med. Wolfram Henn
Prof. Dr. rer. nat. Ursula Klingmüller
Stephan Kruip
Prof. Dr. theol. Andreas Lob-Hüdepohl
Prof. Dr. phil. habil. Annette Riedel
Prof. Dr. iur. Stephan Rixen
Prof. Dr. iur. Dr. phil. Frauke Rostalski
Prof. Dr. theol. Kerstin Schlögl-Flierl
Dr. med. Josef Schuster
Prof. Dr. phil. Mark Schweda
Prof. Dr. phil. Judith Simon
Prof. Dr. phil. Muna Tatari

External expert
Prof. Dr. phil. habil. Dr. phil. h. c. lic. phil. Carl Friedrich Gethmann (member of the Council up to 13 February 2021, thereafter involvement as an external expert)