AI @ Work: Human Empowerment or Disempowerment?

Sabine T. Koeszegi

If you could train an AI to be a Buddhist,
It would probably be pretty good.
Reid Hoffmann

Abstract Recent advancements in generative AI systems fuel expectations that AI will free workers to resolve creative, complex, and rewarding tasks by automating routine and repetitive work. Furthermore, algorithmic decision systems (ADS) will improve decision quality by providing real-time information and insights, analyzing vast amounts of data, and generating recommendations to support decision-making. In this narrative, AI empowers workers to achievements that they could not reach without the technology. However, using AI in work contexts may also lead to changes in workers' roles and identities, leading to feelings of reduced self-efficacy and lower confidence in their abilities and a sense of diminished value in the workplace, their ethical decision-making abilities, and professional integrity. Initial empirical findings on the impact of AI in the work context point to essential design aspects that will determine which of the narratives becomes a reality. This chapter presents these initial findings and makes design suggestions.

This book chapter is adapted from previous work published in Koeszegi, S.T., Zafari, S. & Grabler, R. (2023): The computer says no: How automated decision systems affect workers' role perceptions in sociotechnical systems; in Garcia-Murillo, M. & Renda A. (Eds): Artificial Intelligence at Work: Interconnections and Policy Implications, Edward Elgar Publishing Ltd., and Koeszegi, S. T. (2023). Automated Decision Systems: Why Human Autonomy is at Stake. In Collective Decisions: Theory, Algorithms And Decision Support Systems, Springer (pp. 155–169).

S. T. Koeszegi (✉)
Institute of Management Science, TU Wien, Vienna, Austria
e-mail: sabine.koeszegi@tuwien.ac.at

H. Werthner et al. (eds.), *Introduction to Digital Humanism*,
https://doi.org/10.1007/978-3-031-45304-5_12

1 Introduction

> I am ChatGPT. ... My main goal is to be a useful tool for people looking for information. I strive to provide accurate and helpful answers as best I can be based on my programming and training. ... I have no personal goals or motivations, as I am an artificial intelligence and have no consciousness or emotions. ... Inside me, I analyze the input I receive, break it down into its parts, and use algorithms to generate an answer based on the patterns and relationships I have learned from my training data. ... My training data consists of much text from various sources, such as books, articles, and websites (Chat GPT, https://chat.openai.com/chat, 07.03.23).

You have just read ChatGPT's answer: "Who are you, what are your intentions, and how do you work?" The achievements of this artificial intelligence and similar tools are impressive. They allow routine and repetitive tasks to automate, freeing workers to focus on more complex and creative work. They provide workers with real-time information and insights by analyzing data, generating recommendations to support decision-making, and improving decision quality. They may facilitate communication and collaboration among workers or provide personalized assistance. These systems can serve as digital mentors or coaches, providing guidance, training, and task feedback, helping workers improve their skills and performance. They may also support generating ideas, drafting content, and giving creative suggestions. In this narrative, AI empowers workers to achievements they could not reach without the technology, and experts speak of another significant breakthrough. With the new generative AI systems, the next milestone in the development of artificial intelligence has been reached in augmenting human capabilities.

This new narrative partly contradicts the earlier narrative by MIT's stars Erik Brynjolfsson and Andrew McAfee, who framed the fundamental change of work through AI and automation as a "Race Against the Machine" (2012) in which AI technologies will more and more replace humans. While other chapters of this book discuss potential applications of AI in work processes and answer questions of which tasks and jobs could be replaced by AI (routine cognitive and manual tasks are particularly prone to automation (Autor et al., 2003)), we are interested in the question of how AI systems will change work and impact our understanding of the roles of humans and machines in collaborative work settings. Using AI in work contexts may lead to changes in workers' roles and identities. As AI automates tasks previously performed by humans, workers may need to adapt to new roles, resulting in self-perception shifts, impacting their self-identity and how they view their role in the workplace. Furthermore, workers may worry about the potential for AI to take over their tasks, leading to feelings of reduced self-efficacy and lower confidence in their abilities and a sense of diminished value in the workplace. Also, workers may feel responsible for the ethical implications of using AI, which may influence their self-perception regarding their ethical decision-making abilities and professional integrity.

At this point, it is too early to assess what impact generative AI will have on people in the work context—which of the two narratives will prevail—whether AI will tend to empower or disempower people. The decision on this will ultimately be

made in the design of AI systems. Initial empirical findings on the impact of AI in the work context point to essential design aspects that will determine which of the narratives becomes a reality. This chapter presents these initial findings and makes design suggestions. In Sect. 2, we will introduce algorithmic decision systems (ADS) and discuss subsequently in Sect. 3 how they impact decision outcomes. Section 4 addresses in detail how ADS will change work, i.e., how tasks are assigned to roles (human or AI); how ADS may affect self-assessment, self-efficacy, and human competencies; and why human oversight and accountability need to be addressed when ADS are at work. Finally, we provide design propositions and conclusions in Sect. 5.

2 Algorithmic Decision Systems

There are countless and incredibly diverse applications of AI. There will be no industry or workplace that will not be affected by partial automation (Brynjolfsson & McAfee, 2012). Routine tasks that generally do not require human intervention (Autor et al., 2003) can be fully automated. At the same time, however, there will be tasks that can only be solved with specific human skills on a cognitive, social, or cultural level. Tasks that require human skills in analytical problem-solving, critical thinking and judgment, creativity, empathy, entrepreneurial skills, leadership, persuasion, or imagination cannot be performed by AI alone. Working conditions will, therefore, increasingly include hybrid work environments where AI systems complement and augment human skills (Daugherty & Wilson, 2018).

This chapter emphasizes cases in which agency is shared between human and machines, i.e., in hybrid activities that require some form of collaboration. The most important tasks that we could transfer to machines are decisions. Automated decision systems (ADS) are "systems that encompass a decision-making model, an algorithm that translates this model into computable code, the data this code uses as an input, and the entire environment surrounding its use" (Chiusi et al., 2020). They are often framed as augmenting AI technology, supporting human decisions and problem-solving processes by enhancing human judgment with machine intelligence-based analytic capabilities.

Informed by the Aristotelian view of phronesis, human judgment includes the following elements (Koutsikouri et al., 2023):

1. *Not knowing*, i.e., considering that the situation also contains unknown dimensions and relates to answering questions like "Where does a problem begin/end?" "What is at stake?" "What is relevant?"
2. *Emotions*: as sensory perceptions, they inform us of what is essential and alert us to something that requires our attention and provide a motivation compelling us to act.
3. *Sensory perception*, which is not reduced to collecting data but is intertwined with meaning and emotions and is part of human sense-making, making them open for sensory impressions despite the influence of prior knowledge and prejudice.
4. *Lived experience*, i.e., cultivated professional knowledge, which paves the way for dealing with the horizons of not knowing.

5. *Intuition*: understood as contextual and embodied (tacit) knowledge that denotes the unconscious knowledge process of pattern recognition through accumulated experiences.
6. *Emisteme* (scientific knowledge) and *techne* (lived experience): reflects the phronetic use of general knowledge and lived experience, which entails knowing when and how to apply rules and principles in a specific situation.

Human judgment is closely tied to action and has a strong collective quality in professional contexts. It can be summarized as a "synthesizing capacity in human action" (Koutsikouri et al., 2023, p. 5298), and algorithmic decision systems cannot replace human judgment. Other than human judgment, ADS relies on known data, cannot change ultimate (pre-programmed) goals, and is disconnected from sense-making and emotions for human-centered decisions. However, it can process immense amounts of data to detect patterns and use the knowledge represented in data available for specific purposes (Dragicevic et al., 2020). Hence, combining these complementary capabilities should empower and enhance human problem-solving capabilities (Agrawale et al., 2019; Krüger et al., 2017).

The idea of being supported by support systems in decision-making emerged in the 1960s. Since then, researchers have developed data and model-based systems to support complex and challenging decision-making (e.g., Kersten & Lai, 2007). However, with data-driven AI methods, the field of application has expanded to simple, ordinary everyday decisions, where we are either supported by pre-selecting suitable alternatives or they make and execute decisions entirely for us (Koeszegi, 2021).

The paradigmatic change of ADS is based on the ever-increasing autonomy and the resulting agency of such systems. Decisions we made ourselves in the past are wholly or partially transferred to ADS. In many applications of algorithmic decision-making, the boundaries between automated decision-making and decision-making support are blurred, and often, humans are unaware that ADS are working in the background.

Applications of algorithmic decision systems are manifold and diverse (see Fig. 1), as are the reasons for using them and their effects on decision quality. For example, recommender systems help in the pre-selection of decision alternatives (e.g., search engines), pattern recognition systems reduce complexity (e.g., medical diagnostics), predictive analytics systems minimize uncertainty and risk (e.g., prediction of creditworthiness), and assistive systems can be used to reduce human errors of judgment (e.g., automatic brake assistants). Hence, ADS are associated with increased efficiency in decision-making, including lower costs and better outcomes (e.g., Smith et al., 2010; Wihlborg et al., 2016).

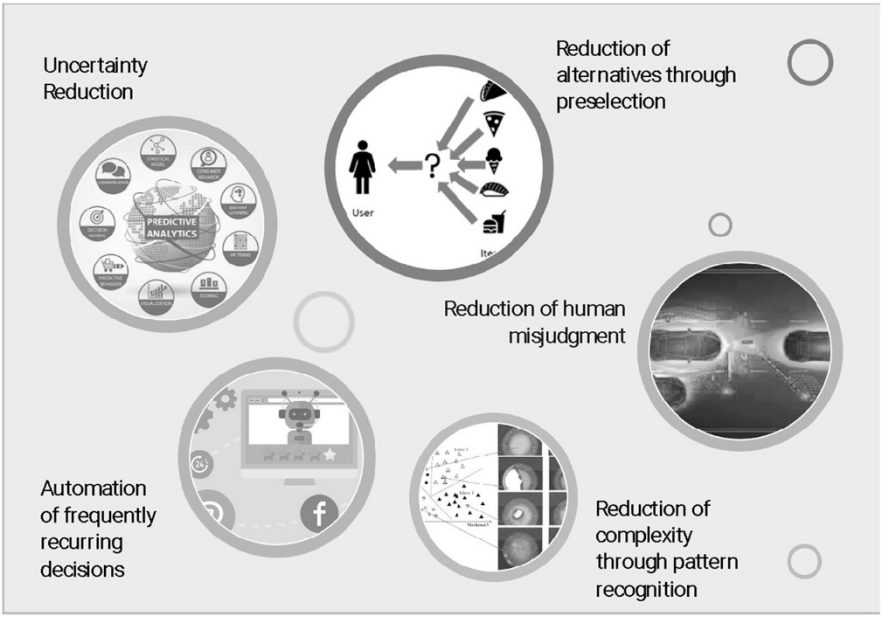

Fig. 1 Applications of algorithmic decision systems

3 How ADS Impact Decision Outcomes

Indeed, under laboratory conditions, combining humans and ADS's complementary capabilities improves decision quality. For instance, human-AI collaboration outperforms human-only or AI-only decisions in diagnosing cancer (Wang et al., 2016). The resulting collaborative success is attributed to the unique advantages that emerge from combining human and AI capabilities in a compatible way (Krüger et al., 2017). Furthermore, a well-designed ADS enhances data analysis by promoting the understanding of multimodal information extracted from multiple data channels, e.g., sorting, scoring, or categorizing the data. At the same time, it reduces the cognitive workload demand for the decision-maker, resulting in improved decision quality (Dragicevic et al., 2018).

ADS is also seen as a game changer in the public sector. Through ADS support, politicians and public servants expect higher efficiency, better service, and higher engagement and professionalism. Ranerup and Henriksen (2019) empirical findings reveal that in the Swedish administration, the new technology in some respects has increased, in association with a focus on citizen-centricity—accountability, decreased costs, and enhanced efficiency. However, they also critically address aspects of negotiating the trade-offs between professional knowledge vs. automated treatment, a potential decrease in costs vs. the increase in service quality, and citizen trust vs. the lack of transparency. Kuziemski and Misuraca (2020) argue that the "public sector's predicament is a tragic double bind:

its obligations to protect citizens from potential algorithmic harms are at odds with the temptation to increase its efficiency—or in other words—to govern algorithms while governing by algorithms."

Overall, people's attitudes toward ADSs seem overly optimistic: Decisions taken automatically by AI are often evaluated on par or better than by human experts (Araujo et al., 2018) even though research that has focused on the effects of algorithmic decisions on those affected has already tempered the high expectations. Indeed, Whittaker et al. (2018, p. 42) conclude in the AI NOW 2018 Report that the harms and biases in AI systems are beyond question: "That debate has been settled, the evidence has mounted beyond doubt." It turns out that algorithmic choices come with severe problems due to partial or incomplete data, inadequate modeling, and problematic objectives and fail with severe consequences (e.g., Citron, 2007; Jackson & Marx, 2017; Murray, 2015; Feijo, 2018; Loewus, 2017; Charette, 2018). Whittaker et al. (2018) cite a string of high-profile examples to show how AI systems perpetuate and amplify social injustice and inequality. Furthermore, in exceptionally high-risk applications of AI systems, such as in healthcare, scholars raise concerns about the prevalence of poor reporting in deep learning studies that assess the diagnostic performance of AI systems to be equivalent to that of healthcare professionals and criticize that "despite the accelerating advance of AI adoption, there has been little high-quality evidence establishing the efficacy of these tools in practice" (Burger, 2022). Also, Bogen and Rieke (2018) list numerous examples of how ADS can perpetuate interpersonal, institutional, and systemic biases due to discrimination based on gender, race, age, or religion. The Berkeley Haas Center for Equity, Gender, and Leadership recently analyzed 133 systems across industries from 1988 and 2021 and found that an alarmingly high share of 44.2% of the systems demonstrates gender bias. Around a quarter of the system has gender and racial bias (Smith & Rustagi, 2021).

ADS's autonomous, complex, and scalable nature introduces ethical challenges and may exacerbate societal tensions and inequalities (Mökander et al., 2021). These features pose different challenges to be resolved, i.e., the system's autonomy makes it hard to assign accountability for failures for outcomes, complexity and opacity of ADS impede to link outcomes (effects) with causes, and scalability implies challenges in managing such systems.

Achieving better decision quality with ADS requires a well-designed human-ADS interface with careful consideration of the more extensive sociotechnical system, i.e., the implementation context. Thus, fully realizing the positive potential of ADS in work processes requires detailed sociotechnical system analysis and design (Zafari et al., 2021). People form (correct or incorrect) expectations about a system's capabilities and assumptions about its reliability and trustworthiness. The design of AI systems has to ensure that there is neither overconfidence in algorithmic decisions nor rejection of superior yet imperfect algorithmic decisions (e.g., Burton et al., 2019). Workers will adjust their roles and self-image in the collaborative work process accordingly. These adaptation processes within such a socio-technical system may jeopardize a clear assignment of tasks and responsibilities and pose additional and novel challenges for work design (Zafari & Koeszegi, 2018). When

decisions are made in a collaborative process, the assessment of decision quality also becomes a key concern, where criteria such as precision and accuracy are far from sufficient. All these aspects require the consideration of social psychological issues in the design of AI that go beyond a human-computer-interaction perspective. In the last decade, the paradigm that technology is shaped by and simultaneously influences the evolution of social structures (Orlikowski, 2007; Zammuto et al., 2007) has become increasingly prevalent. In the following, these aspects will be discussed in more detail.

4 How ADSs Change Work

To date, little attention has been paid to analyzing the impact of ADS on human actors within a socio-technical system. Anecdotal evidence from early experiences with factory robotization refers to applauding workers during robot failures. It made the workers feel better about themselves because robots also failed and were not perfect. Following the narrative that imperfect humans are being replaced by flawless, intelligent, precise, and efficient machines, this emotional response from workers is only understandable. It can be expected that using ADS will have lasting effects on people's self-perception and self-efficacy. In the following, we discuss three aspects—addressed by Bainbridge in 1986 as ironies of automation—that will inevitably change as a result of the shift of decisions and agency from humans to ADSs (see Fig. 2):

1. Usually, only those tasks are automated, which can be easily automated, rather than those that should be automated (e.g., because they are stressful, unhealthy, monotonous, etc.). These design errors prevent the realization of an optimal synergy between humans and machines.
2. Delegating tasks to AI systems can also negatively impact human competencies and know-how long-term. Meaningful experiences are no longer gained; essential skills and abilities are only recovered if needed and trained. In addition, in the

Fig. 2 Challenges in human-ADS collaboration

long run—without immediate feedback on the quality of a decision—a flawed self-assessment may result, either in a bias against automation or in complacency. Again, in the long run, these dynamics will negate the potential benefits of AI decision support.

3. Finally, despite all the automation, humans still need to be tasked with monitoring AI systems and taking responsibility for the process and outcome—a task humans cannot solve in the case of opaque ADS. This leads to organizational and legal challenges regarding assigning responsibility and liability.

4.1 Assignment of Tasks to Roles

ADS significantly impact work processes, individuals' tasks, and their understanding of their roles. The British TV comedy show *Little Britain* presents these changes satirically. In this sketch, a mother visits the hospital with her 5-year-old daughter to make an appointment for a tonsil operation. After the receptionist enters the daughter's details, she says the child is scheduled for bilateral hip surgery. Despite the mother's objections, which the receptionist initially types into her computer, she repeatedly responds with the answer: "The computer says no!" Regardless of how reasonable the mother's objections and how wrong the computer's statements are, the machine's suggested decision ultimately prevails. The satire reveals how supposedly "intelligent systems" can absurdly shift the roles and responsibilities of humans and machines and that "clever" devices with their mathematical algorithms are trusted too much.

Practical experience and scientific studies confirm this satire to be more realistic than we know and draw attention to critical aspects (Fig. 3).

In a case study of a Swedish government agency, where an ADS is used to assess the eligibility of applicants for government benefits, the shift in role structure is

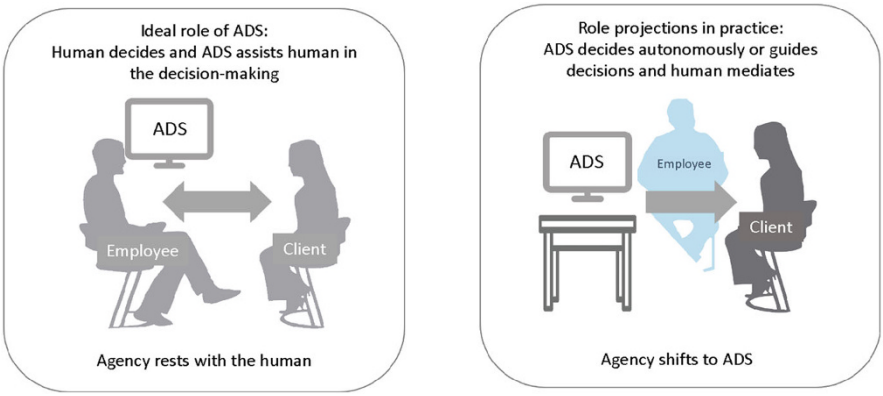

Fig. 3 Role projections and agency in practice

visible (Wihlborg et al., 2016). Whereas previously, the staff members made assessments and decisions based on the application material, they increasingly see themselves only as mediators between the system and the applicants. They "just keep the system running," although they are formally responsible for the final decision. Officials point out in interviews that the system proposes a decision based on all the information entered; therefore, there can be no room for doubt about the decision. They are convinced that the ADS cannot be wrong and attribute high competencies to the system.

In contrast, the perception of their competencies and the capacity to act (in comparison) is perceived as lower. Hence, the assignment of tasks is associated with corresponding expectations and attributions of competence, while one's ability to act is equally restricted. Accordingly, Wihlborg et al. (2016) highlight how decision-support users become mediators rather than decision-makers. While self-determined action requires a degree of personal accountability, delegating decisions to automated systems limits this agency and perceived control over the decision-making process.

This role shift is a consequence of most digitalization strategies, which focus mainly on enhancing machine intelligence and industrial productivity by considering workers as mere "users" rather than collaborators of these systems. When ADS restrict human roles to exercise or communicate ADS decisions, it becomes increasingly difficult for humans to assume accountability for the whole process and outcomes or to take corrective action in the case of a system failure or system errors.

How the user's perception of the potential roles of ADS are entangled with the self-assessment of their competence shows in an experimental study by Jacobsen et al. (2020) in Denmark. In a collaborative waste sorting task for recycling, laypersons received classifications of items and confidence scores by an AI system. At the same time, participants performed better with ADS overall, people who were supported perceived themselves as less effective than they were (underconfident). In contrast, the opposite happened for people with AI support (overconfident). In a qualitative analysis of post-experimental perceptions of the participants of the same study, Papachristos et al. (2021) identified four distinct roles as projections of what users expected from the interaction with the ADS: People who self-assessed relatively high competence in waste sorting (which is a non-trivial task) expected the system to confirm their decision and would ignore a system's opposite suggestion unless they were very unsure. Here ADS were assigned a mirror role (inspired by the fairy tale of Snow White) or an assistant role. In both cases, the agency for the decision resided in the human. Participants who self-assessed, in contrast modest to low, trusted the AI system to have a better judgment. Here, the system was assigned a guiding or even an oracle role, shifting the agency to the system.

It is critical to understand that self-perception is determined by whether users of ADS receive feedback about the correctness of their ultimate decision. Papachristos et al. (2021) find, indeed, if a task that is conducted over a significant period happens to be undertaken with a wrongful decision, still, the user always felt competent; this could impact the overall performance of the collaborative achievement and cease a potential positive impact of ADS in decision-making. On the other hand, if people do

not get feedback about the correctness of their decisions, over time, this could decrease their competence and overconfidence in ADS. In the next section, we discuss in more detail how using ADS may impact skills, competencies, and self-assessment.

To rely on algorithmic decisions, human decision-makers must feel in control (Dietvorst et al., 2016; Meurisch et al., 2020). When individuals experience control over work processes and outcomes, they also feel comfortable collaborating with proactive and autonomous AI systems (Zafari & Koeszegi, 2020). Hence, providing working conditions that preserve a sense of control and efficacy is vital.

4.2 Self-Assessment, Self-Efficacy, and Human Competences

Human decision-makers will have developed expectations of what the specific ADS can do, should do, and how it functions. These expectations may be formed through personal exposure to similar systems and second-hand experiences from coworkers or media. These preexisting expectations will influence how systems are used and relied on (Burton et al., 2019). Zhang et al. (2021) surveyed people's expectations of AI systems. They found that the preferred characteristics of AI systems do not only relate to instrumental capabilities, but they also expect human-like behavior, performance orientation, and shared understanding between humans and these systems. False expectations about the actual capacities of AI systems may lead to either an automation bias, i.e., over-trust in a system, or, to the contrary, a reluctance to rely on AI systems, i.e., a so-called algorithm aversion (Burton et al., 2019).

On the one hand, the mere fact that a decision is made by a machine, not by a human being, lends it a certain degree of legitimacy and neutrality (Citron & Pasquale, 2014). This can simultaneously weaken trust in one's expertise and competence. Empirical studies show that lay people with little or no knowledge of a particular domain prefer to trust an algorithm rather than rely on human expertise. At the same time, experts are significantly less likely to rely on the credibility of algorithmic predictions (Logg et al., 2019). Wouters et al. (2019) also show how laypeople can be impressed even by obviously incorrect ADS outputs: In their study, an ADS with face recognition software identified not only the gender, ethnicity, and age of subjects but also their emotional state and personality traits. Incorrect classifications by the system did not lead to distrust in some subjects but, on the contrary, even caused some subjects to question their self-image: "[The display] must be correct. Because a computer is doing the evaluation, and computers are better than humans at making those kinds of conclusions" (Subject, quoted in Wouters et al., 2019, 454). This automation bias discourages people from questioning system decisions or seeking more information. It leads to inappropriate monitoring of automated functions, i.e., complacency, and creates, in the long run, a human dependency on ADS (Parasuraman & Manzey, 2010; Bahner et al., 2008).

On the other hand, studies also report a so-called Dunning-Kruger effect (DKE) (see He et al., 2023), a metacognitive bias due to which individuals overestimate

their competence and performance compared to algorithmic support. An inflated, false self-assessment and illusory superiority despite poor performance can lead to an under-reliance on AI systems. He et al. (2023) find that a linear relationship cannot explain the interaction between self-assessment and reliance on AI systems. Instead, they suggest that explanations by the system about the flawed nature of AI advice may mitigate the lack of trust and DKE. Thus, to develop an appropriate level of trust in the technology, algorithmic education is needed to create reasonable expectations of ADS and its capabilities: people need to be trained not only in their area of expertise but also in how to interact with algorithmic tools and interpret their results, including teaching important core statistical concepts such as error and uncertainty. Users must be exposed to errors in automation during training to mitigate the risk of complacency, misuse of automation, and bias against automation (Bahner et al., 2008). However, the gravity of decisions also influences human reliance on AI systems. The more serious the consequences of decisions are, the more people are reluctant to rely on algorithms (Filiz et al., 2023). Only when workers understand the decision-making process can they evaluate the consequences of their decisions and gain new knowledge to overcome algorithm aversion (Adadi & Berrada, 2018).

Another critical factor influencing AI system support perception is the timing of the decision support. As discussed earlier, workers using ADS may feel reduced to the role of a recipient of the machine decision and affect user acceptance and be perceived as reputational damage when using such systems, as decision-makers feel they have less opportunity to demonstrate their expertise. This could be avoided by providing decision support after decision-makers have processed information to decide and use the ADS as additional information. Langer et al. (2021) show that these users show a steep increase in self-efficacy in the task and are more satisfied with their decision.

Skills and competencies might deteriorate when they are not used and trained regularly. Hence, deploying ADS may also lead to de-skilling processes of workers. Also, early experiences with automation show that when humans no longer acquire essential expertise or skills—or lose them over time—automated systems replace them (Bainbridge, 1983). At the same time, the skills required of workers also change with the use of ADS. Smith et al. (2010) show in their study how even low-level automation can significantly impact workers' skill levels. For example, the introduction of electronic vote-counting machines turned the previously relatively simple routine task of counting votes by hand into a complex problem about cybersecurity requiring know-how about algorithmic systems and data security. While the switch to an automated system is intended to prevent human error in the counting process, it also creates new challenges—and thus sources of error—because of the need to operate and monitor these systems. Looking at this example, it is still being determined whether the great hopes for efficiency and avoiding human error will be achieved. It seems more like a shift in the potential causes of errors.

Another interesting study analyzes how the use of ADS affects what humans know—and an AI does not know—that is, the unique human knowledge we

described earlier as phronesis. As discussed earlier, ADS are framed as a complementary technology to humans, supporting human decisions and problem-solving processes by enhancing human judgment with machine intelligence-based analytic capabilities. Fügener et al. (2021) analyze the effect of this joint decision-making on the knowledge of humans. They not only look at the individual level but also analyze AI's impact on the "wisdom of crowds." After a set of different controlled experiments, they conclude that humans interacting with artificial intelligence behave like "Borgs," i.e., cyborgs with high individual performance but without human individuality, resulting ultimately in loss of unique human knowledge and leading to long-term adverse outcomes in a variety of human-AI decision-making environments. Their simulation results also suggest that groups of humans interacting with AI are far less effective than those without AI support.

All these results indicate that using ADS does not necessarily always lead to improvement in decision outcomes. On the contrary, long-term adverse effects on self-assessment, self-efficacy, and unique human know-how can cancel out ADS's positive effects and lead to a worse performance of the socio-technical decision than if humans would decide alone.

4.3 Human Oversight and Accountability

Transferring decision-making to ADS includes transferring (part of) the control over the decision-making process and the actual decision from humans to artificial agents while keeping humans accountable for the outcomes of the decisions. Hence, human actors might be accountable for a system's wrong decision. Such accountability without control over the decision-making process creates ethical issues and tension within organizations that need to be considered and addressed before systems are deployed. Furthermore, from a legal perspective, within the EU, some fully automated decisions concerning natural persons are prohibited by Article 22 of the EU General Data Protection Regulation (GDPR 2016), which stipulates that natural persons have the right not to be subjected to a decision based solely on automated processing, including profiling. They have furthermore the right to access meaningful explanations of algorithmic decisions.

Generally, it seems complicated to accept ADS as legitimate if they replace humans in critical decisions (Simmons, 2018). Smith et al. (2010) illustrate this in an example where the use of automated fingerprint identification systems affects the decision-making of experts: the experts' final decision is based on a recommendation for the most likely match of the fingerprint, leaving some experts even unable to explain how the decision was derived as it is beyond their comprehension and scrutiny. According to Smith et al. (2010), this shows two dysfunctions of accountability in ADS-supported decisions: (1) experts are relying more on outputs by the machine while not understanding the decision process, and (2) experts can be blamed for false accusations of a crime as it is them who make the final decision. Nevertheless, they cannot be blamed entirely as the automated system had a part in it

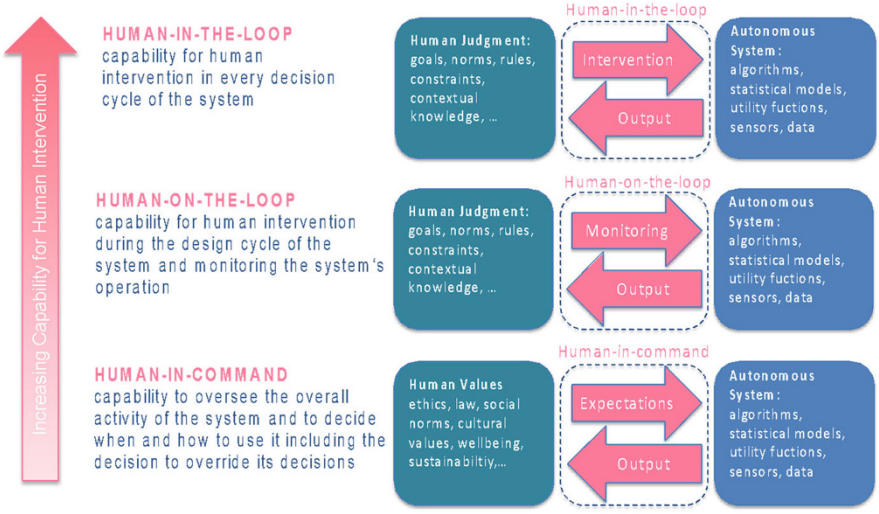

Fig. 4 Human intervention in AI systems

(i.e., diffusion of accountability). Other studies show that as the autonomy of an autonomous AI system increases, people attribute more blame to the system than to themselves (Kim & Hinds, 2006; Furlough et al., 2021). One possible explanation is that people perceive these autonomous agents to have more agency and freedom in deciding and are thus automatically subjected to taking the blame for the choice. However, this is different for taking credits. Lei and Rau (2021) find that autonomous AI systems are more blamed than human agents, but they both received similar levels of positive recognition. Thus, introducing independent agents to work processes challenges traditional accountability practices.

Systems, therefore, should allow different degrees of oversight and control depending on the AI system's application and the potential risk of bad decisions. Figure 4 shows how the concepts of "human-in-the-loop," human-on-the-loop, and human-in-command differ in the degree of human intervention.

The human-in-command approach is only suitable for fast, standardized, and frequent decisions in a well-defined context, where only little negative consequences of an incorrect decision by the AI are to be expected. AI-based technologies cannot meet the requirements for moral agency and accountability (Coeckelbergh, 2019; Zafari & Koeszegi, 2018). Hence, decisions with high uncertainty of outcomes that involve significant ethical issues require human involvement in the form of a human-in-the-loop or human-on-the-loop approach. Decisions that need high transparency should be left entirely to humans (Ivanov, 2022). Tatiana Cutts (2022) questions whether human oversight principles are sufficient to ensure ethical standards for ADS-assisted decision outcomes. The principle of oversight provides that humans should play a corrective role, particularly in critical decisions (such as the relative priority of organ transplants for patients or whether to hire an applicant or fire an

employee or how to sentence a defendant). The assumption is that applying human judgment is both a necessary and sufficient safeguard against unjust decisions. However, safeguarding fundamental rights requires not only human judgment in the decision-making process itself but also gatekeeping, i.e., making principled decisions about the use of ADS only after ensuring that they take into account the right considerations in the right way. Hence, ensuring that workers are able and willing to take responsibility for ADS-supported decisions falls short. The required gatekeeping functions must be assumed by the management, taking overall accountability for ADS implementations.

In addition, ADSs also raise interesting liability issues when the ability of humans to control technical systems is limited (Wagner, 2019). Because most regulatory mechanisms follow a pattern of binary liability by regulating either human or machine agency but are not allowing meaningful liability for socio-technical decision-making, regulatory gray areas arise where human rights are challenged. Although specific regulatory mechanisms exist for purely automated decision-making, they do not apply once humans sign off automated decisions. Wagner (2019) concludes that ADS-based decision-making is quasi-automation, which is only a rubber-stamping mechanism for fully automated decisions.

As described earlier, a lack of human control and oversight can also result from overdependence on ADS. One possible strategy to mitigate this problem is to develop so-called reflection machines (Haselager et al., 2023) that provide meaningful human problem overview and control through their specific form of decision support. A reflection machine does not give the decision-makers suggestions for a decision but instead challenges their reasonings and decisions. Reflection machines, for example, point to facts or raise critical issues or counterarguments inconsistent with the proposed decision to improve the problem-solving process and decision quality. This support also increases people's problem-solving skills and counteracts an unreflective reliance on ADS. Especially in the medical field, experimental studies already show positive experiences in this regard (Haselager et al., 2023).

Whether people are willing to take responsibility for ADS-supported decisions depends mainly on the extent to which they understand the "inner workings" and how an ADS makes decisions. In other words, a lack of transparency about the system design, the system's objectives, and how a decision recommendation is reached (the so-called black-box problem) weaken people's willingness to take responsibility for the decisions. Therefore, ADS should be able to close the knowledge gap, which is also understood as reducing the information asymmetry between the system and the users (Malle et al., 2007). Hence, system transparency increases technology acceptance and establishes appropriate trust (Miller, 2019). Expanding system transparency allows users to understand better the performance of the system and the processes that lead the system to make a particular decision or prediction (Felzmann et al., 2019; de Graaf & Malle, 2017). However, it is not enough to have access to the data processed by the system; the results of ADS must be accompanied by explanations of how and why a decision was made. Users must view the results of ADS as plausible, valuable, and trustworthy (Papagni et al., 2022; Papagni & Koeszegi, 2021a, 2021b). The focus is not on the exactness of the explanation but

on the understanding and plausibility of explanations that result from contextual negotiations between the system and its users. When a system explains its decision-making process in a language that workers can understand, they better understand the causes and premises associated with the decision. In this way, explaining can help workers consider outcomes more thoroughly, manage the problem, and feel accountable for the outcome of the decision process.

5 Conclusions

Recent advancements in generative AI systems fuel expectations that AI will free workers from routine and repetitive work for creative, complex, and rewarding tasks. Furthermore, ADS will improve decision quality by providing real-time information and insights, analyzing vast amounts of data, and generating recommendations to support decision-making. In this narrative, AI empowers workers to achievements they could not reach without the technology. However, using AI in work contexts may also lead to changes in workers' roles and identities, leading to feelings of reduced self-efficacy and lower confidence in their abilities and a sense of diminished value in the workplace, their ethical decision-making abilities, and professional integrity. We argued that whether AI will empower or disempower, people will ultimately depend on the design of AI systems.

Based on this analysis of the first empirical evidence, we conclude that—next to the empowering capacity of ADS—these systems can also enable human error, reduce human control, eliminate human responsibility, and devalue human capabilities. ADS affects our self-image by pushing us from the active role of decision-maker to the passive role of a mediator or facilitator. At their worst, these systems limit our autonomy and undermine human self-determination. The extent to which these potentially detrimental effects of AI systems on workers come to fruition depends to some extent on the system's design and legal regulations of our fundamental rights.

Complementing human decision-making and problem-solving processes inherently requires the integration of two distinct thought processes: that of the human and that of the AI system. Both processes must be mapped and understood transparently enough to create cognitive compatibility. Otherwise, these processes run in parallel, and algorithmic systems combat rather than enhance human decision-making (Burton et al., 2019). In other words, AI systems must be designed to work for humans, not the other way around. This requires understanding human decision-making processes from a rational and normative decision-making perspective and a sense-making perspective that recognizes the context-dependence of decision-making processes (Papagni & Koeszegi, 2021b).

Transforming today's notion of human-AI collaboration into tomorrow's organizational reality requires specific reference models, procedures, standards, and concrete criteria for appropriately considering human factors in the development and implementation of ADS. In other words, creating a sociotechnical system requires

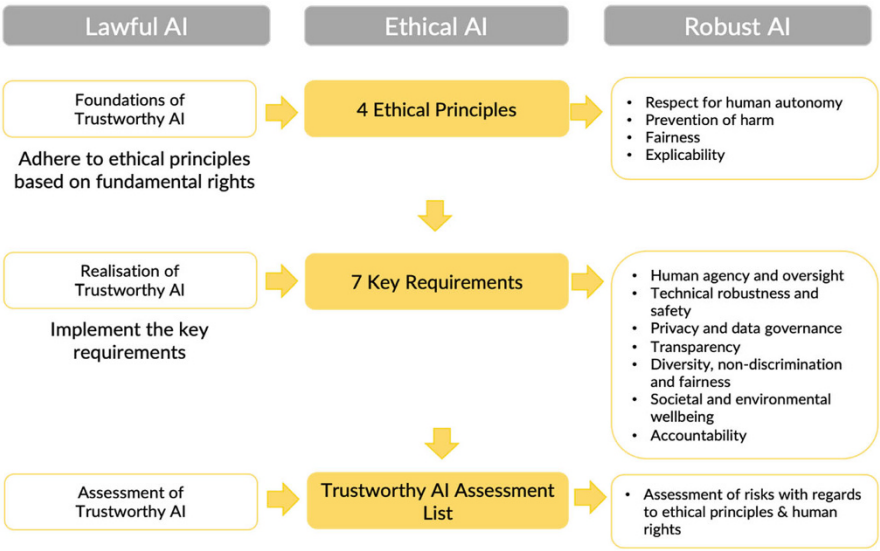

Fig. 5 AI HLEG trustworthy AI

the consideration of both technical and social aspects of work processes to illuminate the mutual influence of technological and social entities. Therefore, it is necessary to identify requirements for human-centered technology design that preserves workers' control and meaningful role. Furthermore, for successful integration of ADS into work organizations, we need to involve workers who use these technologies in their work in the development process rather than presenting them with a fait accompli and requiring them to take responsibility for decisions dictated by ADS, leaving them with only the role of rubberstamping automated decisions. To achieve this, governance mechanisms are needed to help organizations design and deploy ADS ethically while enabling society to reap the full economic and social benefits of automation (Mökander et al., 2021).

Value-based design processes like the IEEE 7000 standard and ethics-based auditing (EBA) as a governance mechanism allow organizations to validate the claims of their systems. Numerous ethics-based frameworks and assessment tools exist for AI systems (see, e.g., Mökander et al., 2021). The framework exhibited in Fig. 5 was developed by the AI High-level Expert Group of the European Commission in 2018 and builds the foundation for the currently negotiated AI regulation proposition within the European member states. It is based on four ethical principles (i.e., respect for human autonomy, prevention of harm, fairness, and explicability). It operationalizes these principles in seven key requirements, which are (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination, and fairness; (6) societal and environmental well-being; and (7) accountability. The Trustworthy AI Assessment List is a proposal for an ethics-based auditing tool that guides system

designers into reflecting questions about potential harm and risks associated with the AI system at hand.

Such ethics-based assessment tools provide a structured process for evaluating AI systems for compliance with relevant ethical principles and human rights (Mökander et al., 2021). If they focus not only on the narrower human-machine interface but also consider the broader socio-technical context of implementation, these tools can effectively realize human empowerment. The trustworthy AI assessment tool considers this broader implementation context through the essential requirement (6) societal and environmental well-being.

We conclude this chapter with exemplary questions from the Assessment List of Trustworthy AI (ALTAI) to inspire discussions in the vein of reflection machines.

This subsection helps self-assess necessary oversight measures through governance mechanisms such as human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approaches.

Please determine whether the AI system (choose the appropriate):

Is a self-learning or autonomous system
Is overseen by a *Human*-in-the-Loop
Is overseen by a *Human*-on-the-Loop
Is overseen by a *Human*-in-Command

Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?
Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?
Did you ensure a "stop button" or procedure to safely abort an operation when needed?
Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?
This subsection helps self-assess the potential effects of your AI system on societal and environmental well-being. The following questions address issues related to work contexts:

Does your AI system impact human work and work arrangements?
What is the potential impact of your system on workers and work arrangements?
Do you ensure that the work impacts of the AI system are well understood?
Did you assess whether there is a risk of de-skilling of the workforce? If there is a risk, which steps have been taken to counteract de-skilling risks?
Did you assess how the system may affect the attribution of capabilities and accountabilities in work contexts?

(continued)

Do you ensure that workers understand how the system operates, which capabilities it has, and which not? If yes, describe measures:

Based on your answers to the previous questions, how would you rate the risk that the AI system negatively impacts work and work arrangements?
How would you rate the measures you have adopted to mitigate this risk?

Discussion Questions for Students and Their Teachers

1. Which risks are associated with the implementation of ADS in work contexts?
2. Which design propositions can mitigate adverse effects of ADS, automation bias, or complacency?
3. Which design propositions have been made to ensure the empowerment of workers rather than disempowerment?

Learning Resources for Students

1. Bainbridge, L. (1983): Ironies of automation. In: Automatica 19 (6), S. 775–779, https://doi.org/10.1016/0005-1098(83)90046-8.
   This paper describes in more detail the three ironies of automation, which are also addressed in this book chapter in Fig. 2.
2. HLEG AI. (2019). High-level expert group on artificial intelligence. Ethics Guidelines for Trustworthy AI. *European Commission.* https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, Accessed 20.04.2023.
   Assessment List of trustworthy AI, European Commission, accessed 23.04.23
   These two deliverables of the AI HLEG of the European Commission build the fundament of the AI regulation act.
3. Chiusi, F. Fischer, S. Kayser-Bril, N. Spielkamp, M. (2020). Automating Society Report 2020. *Algorithm Watch.* https://automatingsociety.algorithmwatch.org. Accessed 20. April 2023.
   This Report of Algorithm Watch comprises an overview of existing AI applications and threats and challenges.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the impact of artificial Intelligence: Prediction versus judgment. *Information Economics and Policy, 47*, 1–6. https://doi.org/10.1016/j.infoecopol.2019.05.001

Araujo, T., De Vreese, C., Helberger, N., Kruikemeier, S., van Weert, J., Bol, N., Oberski, D., Pechenizkiy, M., Schaap, G., & Taylor, L. (2018). Automated decision-making fairness in an AI-driven world: Public perceptions, hopes and concerns. *Digital Communication Methods Lab*. https://hdl.handle.net/11245.1/369fdda8-69f1-4e28-b2c7-ed4ff2f70cf6

Autor, D. H., Levy, F., & Murane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *The Quarterly Journal of Economics, 118*(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Bainbridge, L. (1983). Ironies of automation. *Automatica, 19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

Bahner, J. E., Hüper, A.-D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies, 66*(9), 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001

Bogen, M., & Rieke, A. (2018, December 9). Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn*. Accessed April 20, 2023, from https://apo.org.au/node/210071

Brynjolfsson, E., & McAfee, A. (2012). *Race against the machine: How the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy*. Digital Frontier Press.

Burger, M. (2022). *The risk to population health equity posed by automated decision systems: A narrative review*. arXiv preprint arXiv:2001.06615. https://doi.org/10.48550/arXiv.2001.06615

Burton, J. W., Stein, M.-K., & Jensen, T. B. (2019). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making, 33*(2), 220–239. https://doi.org/10.1002/bdm.2155

Charette, R. N. (2018, January 24). *Michigan's MiDAS unemployment system: Algorithm alchemy created lead, not gold*. IEEE Spectrum. Accessed April 20, 2023, from https://tinyurl.com/6vey252h

Chiusi, F., Fischer, S., Kayser-Bril, N., & Spielkamp, M. (2020). *Automating Society Report 2020*. Algorithm Watch. Accessed April 20, 2023, from https://automatingsociety.algorithmwatch.org

Citron, D. K. (2007). Technological due process. *Washington University Law Review, 85*, 1249.

Citron, D. K., & Pasquale, F. (2014). The scored society: Due process for automated predictions. *Washington University Law Review, 89*, 1.

Coeckelbergh, M. (2019). Artificial intelligence, responsibility attribution, and a relational justification of explainability. *Science and Engineering Ethics*, 1–18. https://doi.org/10.1007/s11948-019-00146-8

Cutts, T. (2022). *Supervising automated decisions*. SSRN Scholarly Paper Nr. 4215108. https://doi.org/10.2139/ssrn.4215108

Daugherty, P. R., & Wilson, H. J. (2018). *Human+ machine: Reimagining work in the age of AI*. Harvard Business Press.

De Graaf, M. M., & Malle, B. F. (2017). *How people explain action (and autonomous intelligent systems should too)*. AAAI Fall Symposia.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming algorithm aversion: People will use imperfect algorithms if the can (even slightly) modify them. *Management Science, 64*(3), 1144–1170. https://doi.org/10.1287/mnsc.2016.2643

Dragicevic, N., Ullrich, A., Tsui, E., & Gronau, N. (2018). A conceptual model of knowledge dynamics in the industry 4.0 intelligent grid scenario. *Knowledge Management Research & Practice, 18*(2), 199–213. https://doi.org/10.1080/14778238.2019.1633893

Dragicevic, N., Ullrich, A., Tsui, E., & Gronau, N. (2020). A conceptual model of knowledge dynamics in the industry 4.0 intelligent grid scenario. *Knowledge Management Research & Practice, 18*(2), 199–213. https://doi.org/10.1080/14778238.2019.1633893

European Commission. *Assessment List of trustworthy Artificial Intelligence (ALTAI) for self-assessment*. Accessed April 23, 2023, from https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

Feijo, S. (2018, July 16). *Here's what happened when Boston tried to assign students good schools close to home*. Northeastern Global News. Accessed April 20, 2023, from https://tinyurl.com/yp5neuxn

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamo-Larrieux, A. (2019). Robots and transparency: The multiple dimensions of transparency in the context of robot technologies. *IEEE Robotics & Automation Magazine, 26*(2), 71–78. https://doi.org/10.1109/MRA.2019.2904644

Filiz, I., Judek, J. R., Lorenz, M., & Spiwoks, M. (2023). The extent of algorithm aversion in decision-making situations with varying gravity. *PLoS ONE, 18*(2), e0278751. https://doi.org/10.1371/journal.pone.0278751

Furlough, C., Stokes, T., & Gillan, D. J. (2021). Attributing blame to robots: I. The influence of robot autonomy. *Human Factors, 63*(4), 592–602. https://doi.org/10.1177/0018720819880641

Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will humans-in-the-loop become Borgs? Merits and pitfalls of working with AI. *Management information Systems Quarterly (MISQ), 45*(3). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3879937

Haselager, P., Schraffenberger, H., Thill, S., Fischer, S., Lanillos, P., van de Groes, S., & van Hooff, M. (2023). Reflection machines: Supporting effective human oversight over medical decision support systems. *Cambridge Quarterly of Healthcare Ethics*, 1–10. https://doi.org/10.1017/S0963180122000718

He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 113*, 1–18. https://doi.org/10.1145/3544548.3581025

HLEG AI. (2019). *Ethics guidelines for trustworthy AI*. Accessed April 20, 2023, from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Ivanov, S. H. (2022). Automated decision-making. *Foresight, 25*(1), 4–19. https://doi.org/10.1108/FS-09-2021-0183

Jackson, D., & Marx, G. (2017, December 6). Data mining program designed to predict child abuse proves unreliable, DCFS says. *Chicago Tribune*. https://tinyurl.com/4wb7yxub

Jacobsen, R. M., Johansen, P. S., Bysted, L. B. L., & Skov, M. B. (2020). Waste wizard: Exploring waste sorting using AI in public spaces. Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society, 1–11. https://doi.org/10.1145/3419249.3420180

Kersten, G. E., & Lai, H. (2007). Negotiation support and e-negotiation systems: An overview. *Group Decision and Negotiation, 16*(6), 553–586. https://doi.org/10.1007/s10726-007-9095-5

Kim, T., & Hinds, P. (2006). Whom should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006 – The 15th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 80–85). https://doi.org/10.1109/ROMAN.2006.314398

Köszegi, S. T. (2021). Automated decision systems: Why human autonomy is at stake. In *Collective decisions: Theory, algorithms and decision support systems* (pp. 155–169). Springer Nature Switzerland AG. http://hdl.handle.net/20.500.12708/30729

Koutsikouri, D., Hylving, L., Lindberg, S., & Bornemark, J. (2023). Seven elements of phronesis: A framework for understanding judgment in relation to automated decision-making. 56th Hawaii Conference on System Sciences (HICSS). https://hdl.handle.net/10125/103280

Krüger, M., Wiebel, C. B., & Wersing, H. (2017). From tools towards cooperative assistants. In *Proceedings of the 5th International Conference on Human Agent Interaction* (pp. 287–294). https://doi.org/10.1145/3125739.3125753

Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy, 44*(6), 101976. https://doi.org/10.1016/j.telpol.2020.101976

Langer, M., König, C. J., & Busch, V. (2021). Changing the means of managerial work: Effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology, 36*(5), 751–769. https://doi.org/10.1007/s10869-020-09711-6

Lei, X., & Rau, P. L. P. (2021). Should I blame the human or the robot? Attribution within a human-robot group. *International Journal of Social Robotics, 13*(2), 363–377. https://doi.org/10.1007/s12369-020-00645-w

Loewus, L. (2017, October 26). Houston District settles lawsuit with teachers' union over value-added scores. *Education Week*. Accessed June 01, 2023, from https://tinyurl.com/yckucffc

Logg, J. M., Minsona, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 15*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *The Journal of Personality and Social Psychology, 93*(4), 491. https://doi.org/10.1037/0022-3514.93.4.491

Meurisch, C., Mihale-Wilson, C. A., Hawlitschek, A., Giger, F., Müller, F., Hinz, O., & Mühlhäuser, M. (2020). Exploring user expectations of proactive AI systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4*(4), 1–22. https://doi.org/10.1145/3432193

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope and limitations. *Science and Engineering Ethics, 27*(4), 44. https://doi.org/10.1007/s11948-021-00319-4

Murray, D. (2015, March 20). Queensland authorities confirm 'miscode' affects DNA evidence in criminal cases. *The Courier Mail*. Accessed June 01, 2023, from https://tinyurl.com/mrxkarpw

Orlikowski, W. J. (2007). Sociomaterial practices: Exploring technology at work. *Organisation Studies, 28*(9), 1435–1448. https://doi.org/10.1177/0170840607081138

Papachristos, E., Skov Johansen, P., Møberg Jacobsen, R., Bjørn Leer Bysted, L., & Skov, M. B. (2021). How do people perceive the role of AI in human-AI collaboration to solve everyday tasks? In *CHI Greece 2021: 1st International Conference of the ACM Greek SIGCHI Chapter* (pp. 1–6). https://doi.org/10.1145/3489410.3489420

Papagni, G. J., De Pagter, J., Zafari, S., Filzmoser, M., & Koeszegi, S. T. (2022). May I explain? Explainability is a Trust Support Strategy for Artificial Agents. Accepted in a special Issue AI4P, AI & Society. *Journal of Knowledge, Culture, and Communication*, 1–14.

Papagni, G., & Koeszegi, S. T. (2021a). A pragmatic approach to the intentional stance: Semantic, empirical and ethical considerations for the design of artificial agents. *Minds & Machines, 31*, 505–534. https://doi.org/10.1007/s11023-021-09567-6

Papagni, G., & Koeszegi, S. T. (2021b). Understandable and trustworthy explainable robots: A sensemaking perspective. *Paladyn, Journal of Behavioral Robotics, 12*(1). https://doi.org/10.1515/pjbr-2021-0002

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors, 52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Ranerup, A., & Henriksen, H. Z. (2019). Value positions viewed through the lens of automated decision-making: The case of social services. *Government Information Quarterly, 36*(4), 101377. https://doi.org/10.1016/j.giq.2019.05.004

Simmons, R. (2018). Big data, machine judges, and the criminal justice system's legitimacy. *UC Davis Law Review, 52*, 1067. https://doi.org/10.2139/ssrn.3156510

Smith, G., & Rustagi, I. (2021). When good algorithms go sexist: Why and how to advance AI gender equity. *Stanford Social Innovation Review*. https://doi.org/10.48558/A179-B138

Smith, M. L., Noorman, M. E., & Martin, A. K. (2010). Automating the public sector and organising accountabilities. *Communications of the Association for Information Systems, 26*(1), 1. https://doi.org/10.17705/1CAIS.02601

Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet, 11*(1), 104–122. https://doi.org/10.1002/poi3.198

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718. https://doi.org/10.48550/arXiv.1606.05718

Whittaker, D., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., Myers West, S., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI now report 2018. AI now institute*.

Accessed April 20, 2023, from https://ec.europa.eu/futurium/en/system/files/ged/ai_now_2018_report.pdf

Wihlborg, E., Larsson, H., & Hedström, K. (2016). *The computer says no!—A case study on automated decision-making in public authorities* (pp. 2903–2912). Örebro University Publications. https://urn.kb.se/resolve?urn=urn:nbn:se:oru:diva-48440

Wouters, N., Kelly, R., Velloso, E., Wolf, K., Ferdous, H. S., Newn, J., Joukhadar, Z., & Vetere, F. (2019). Biometric mirror: Exploring values and attitudes towards facial analysis and automated decision-making. *Conference on Designing Interactive Systems, 1145*. https://doi.org/10.1145/3322276.3322304

Zafari, S., & Koeszegi, S. T. (2018). Machine agency in socio-technical systems: A typology of autonomous artificial agents. In *2018 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)* (pp. 125–130). doi:https://doi.org/10.1109/ARSO.2018.8625765.

Zafari, S., & Koeszegi, S. T. (2020). Attitudes toward attributed agency: Role of perceived control. *International Journal of Social Robotics*, 1–10. https://doi.org/10.1007/s12369-020-00672-7

Zafari, S., Köszegi, S. T., & Filzmoser, M. (2021). Human adaption in the collaboration with artificial agents. In J. Fritz & N. Tomaschek (Eds.), *Konnektivität Über die Bedeutung von Zusammenarbeit in der virtuellen Welt* (pp. 97–106). Waxmann Verlag GmbH. http://hdl.handle.net/20.500.12708/30581

Zammuto, R. F., Griffith, T. L., Majchrzak, A., Dougherty, D. J., & Faraj, S. (2007). Information technology and the changing fabric of the organisation. *Organization Science, 18*(5), 749–762. https://doi.org/10.1287/orsc.1070.0307

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). "An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW3), 1–25. https://doi.org/10.1145/3432945