On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives

Sarah Sterz
Dependable Systems and Software,
Saarland University
Saarland Informatics Campus,
Saarbrücken, Germany
sterz@depend.uni-saarland.de

Kevin Baum
Neuro-Mechanistic Modeling,
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
kevin.baum@dfki.de

Sebastian Biewer
Dependable Systems and Software,
Saarland University
Saarland Informatics Campus,
Saarbrücken, Germany
biewer@depend.uni-saarland.de

Holger Hermanns
Dependable Systems and Software,
Saarland University
Saarland Informatics Campus,
Saarbrücken, Germany
hermanns@cs.uni-saarland.de

Anne Lauber-Rönsberg
IRGET, Faculty of Humanities and
Social Science, TU Dresden
Dresden, Germany
anne.lauber-roensberg@tu-dresden.de

Philip Meinel
IRGET, Faculty of Humanities and
Social Science, TU Dresden
Dresden, Germany
philip.meinel@tu-dresden.de

Markus Langer
Department of Psychology,
University of Freiburg
Freiburg, Germany
markus.langer@psychologie.uni-freiburg.de

ABSTRACT

Human oversight is currently discussed as a potential safeguard to counter some of the negative aspects of high-risk AI applications. This prompts a critical examination of the role and conditions necessary for what is prominently termed *effective* or *meaningful* human oversight of these systems. This paper investigates effective human oversight by synthesizing insights from psychological, legal, philosophical, and technical domains. Based on the claim that the main objective of human oversight is risk mitigation, we propose a viable understanding of effectiveness in human oversight: for human oversight to be effective, the oversight person has to have (a) sufficient causal power with regard to the system and its effects, (b) suitable epistemic access to relevant aspects of the situation, (c) self-control, and (d) fitting intentions for their role. Furthermore, we argue that this is equivalent to saying that an oversight person is effective if and only if they are morally responsible and have fitting intentions. Against this backdrop, we suggest facilitators and inhibitors of effectiveness in human oversight when striving for practical applicability. We discuss factors in three domains, namely, the technical design of the system, individual factors of oversight persons, and the environmental circumstances in which they operate. Finally, this paper scrutinizes the upcoming AI Act of the European Union – in particular Article 14 on Human Oversight – as an exemplary regulatory framework in which we study the practicality of our understanding of effective human oversight. By analyzing the provisions and implications of the European AI Act proposal, we pinpoint how far that proposal aligns with our analyses regarding effective human oversight as well as how it might get enriched by our conceptual understanding of effectiveness in human oversight.

CCS CONCEPTS

• **Social and professional topics** → **Governmental regulations**; **Computing / technology policy**; • **Human-centered computing**; • **Computing methodologies** → *Philosophical/theoretical foundations of artificial intelligence*;

KEYWORDS

Human Oversight, AI Act, High-Risk AI, Law, Psychology

ACM Reference Format:
Sarah Sterz, Kevin Baum, Sebastian Biewer, Holger Hermanns, Anne Lauber-Rönsberg, Philip Meinel, and Markus Langer. 2024. On the Quest for Effectiveness in Human Oversight: Interdisciplinary Perspectives. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3630106.3659051

We want to thank Sven Hetmank and Franz Lehr for their contributions to previous work that this paper is based on. Also, we want to thank three anonymous reviewers for their valuable feedback.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3659051

1 INTRODUCTION

Effective or meaningful[1] human oversight is a notorious centerpiece of ethical guidelines [41] and global legislation [22, 31] intended to govern the deployment of AI-based systems, particularly in high-risk contexts. The regulatory significance of human involvement in automated decision-making is unmistakable, as evidenced by a range of legislation [19, 31], including the overarching principles embedded in the General Data Protection Regulation of the EU [24] and more targeted legislation, such as the European AI Act [22], which mandates human oversight for high-risk AI applications. Research has made strides in defining and assigning responsibility for human oversight [19], or placing it within a broader risk management framework [31]. Human oversight has also been subject to criticism, with concerns being raised about the efficacy of oversight in practice and its potential role as a mere vehicle to legitimize imperfect AI systems [31, 89].

Effective oversight has been discussed by many [19, 31, 49, 55]; however, a question that is crucial for this discussion has not been sufficiently addressed: *When is human oversight effective?* The pressing need for conceptual clarity in this regard is evident, particularly given that the European Union's AI Act is due to be adopted in 2024. Nevertheless, the definition and conditions for achieving *effective* human involvement remain underexplored, highlighting a critical gap in the current state of the art of AI governance.

The following core contributions are made by this paper, and as such provide a structural frame for it:

- We suggest a viable understanding of effectiveness in the context of human oversight based on the idea that the central objective of oversight is risk mitigation. Namely, we propose that an oversight person is effective if and only if they have (a) sufficient causal power with regards to the system and its effects, (b) suitable epistemic access to relevant aspects of the situation, (c) self-control over their own actions, and (d) fitting intentions for their role. We point out that (a)–(c) are jointly sufficient for moral responsibility.
  So, phrased boldly, we could say that

  Effectiveness = Moral Responsibility + Fitting Intentions.
- We exemplify facilitators and inhibitors of effectiveness in human oversight in three categories: the technical design of the system, the individual factors of the human who is in charge of oversight, and the environmental circumstances in which they operate.
- As a litmus test, we look at the details of the European AI Act relative to our proposed conceptualization. We identify possible benefits and shortcomings of the AI Act. In particular, we argue that our conceptualization provides both a more general *and* a more practically useful conceptualization of effective human oversight compared to the respective stipulations in the AI Act.

2 RELATED WORK

Researchers have spelled out who will be in charge of (ensuring) human oversight and have analyzed how human oversight fits into a more global risk management concept (as, e.g., in the AI Act [22]) together with other approaches on accurate and robust systems, data-privacy adhering systems, and transparent systems [19, 49]. The concept of human oversight, however, also faces substantial criticism, as empirical evidence suggests that humans may be ineffective in supervising AI-based systems [31]. This has prompted a call for a more institutional oversight approach, emphasizing the need for empirical evidence regarding humans' capability to oversee AI systems prior to enforcing the oversight [31]. This section will explore the literature on effective oversight from different disciplines in more detail.

Effective human oversight in legislation and guidelines. Oversight is an essential aspect of various ethical guidelines and legislation on the use of AI-based systems in high-risk contexts. After the analysis of 41 policy documents around the globe, Green [31] observes a call for limiting solely automated decisions (e.g., Article 22 of the European General Data Protection Regulation [24]), and an emphasis on the need for human discretion in AI-based decision-making (e.g., Canadian Directive on Automated Decision-Making [30]). The requirement for effective oversight is particularly highlighted in the European AI Act. At the time of writing, the AI Act has not yet been formally adopted. However, a final version was approved by the EU Parliament in March 2024 [23] and amended as part of the so-called corrigendum procedure [22]. We will therefore use this amended version but will refer to earlier versions when relevant.

Human-centric AI. The focus on human oversight seems to stem from the background of ensuring human-centric development and deployment of AI-based systems. One of the most prominent concerns regarding the wide-range deployment of AI-based systems seems to be "that technological development and rationalized efficiency will take place at the cost of human agency and safety or rights" [19]. An obvious way to counter such a techno-centric deployment of AI would be a human-centric development and deployment of AI-based systems. As Enqvist [19] notes: "Human centricity, as it has come to be (broadly) understood, does not only reflect that human needs are to be met by new technologies, but also incorporates the aim to safeguard individual rights and increase human well-being." Indeed, in computer science, a popular area of research is concerned with *human-centered AI* (HCAI) where efficient collaboration between humans and AI systems is investigated [3, 11, 17, 46, 71, 81].

Objectives of effective human oversight. One of the foundational assumptions for the effectiveness of oversight measures is that human oversight should not be an end in itself [49], but that the involvement of humans in contexts where AI-based systems affect high-risk decision-making can make things better in some regard. For example, one hope is that humans may be better than AI-based systems at incorporating ethical considerations and social norms into decision-making contexts [19]. Another hope is that humans might be better at judging single or unusual cases, for instance, in decisions that would affect individual human beings [7, 51]. Additionally, associating the term "effective" with human oversight

[1] Henceforth, we focus exclusively on the term "effective" and omit the term "meaningful". Given their synonymous use in the context of human oversight, any statements regarding effectiveness also extend to meaningfulness. Furthermore, we are only interested in *human* oversight, so we occasionally drop the "human" and talk about "oversight" only for the sake of simpler wording. In response to a very insightful reviewer comment, we have also avoided the phrase "human overseer" because of its problematic connotation in US-American contexts.

shows the hope that a human can effectively contribute to a joint decision situation together with an AI-based system, thus enabling decisions better than those of either the human or the system alone [4, 11, 71]. Other hopes include increased safety [45], improved accuracy [42], or better trustworthiness and more actual trust [55]. The most important objective of human oversight recently seems to be the mitigation of risks to fundamental rights, both directly (as one reading of the AI Act [22] suggests), or indirectly via, e.g., the goal to improve safety or the avoidance of ethically and socially undesirable outcomes [19].

Criticism of human oversight. However, these foundational assumptions have received substantial criticism, calling into question the usefulness of human oversight [31, 89]. Critics (in particular [31]) point to a lack of empirical evidence showing that we, as of now, can reliably integrate human and system abilities in a way that leads to better decisions compared to decisions that every single entity would produce in isolation [4, 5, 65]. For instance, there seem to be substantial challenges with respect to achieving adequate trust in AI-based systems [3, 37, 50, 90]. Humans may overly rely on the outputs produced by automated systems, leading to situations where they do not detect erroneous or unfair outputs [62]. In other situations, humans may have too little trust in AI-based systems or too high confidence in their own abilities, which leads to people overriding actually accurate or fair system outputs [32]. Especially in high-reliability contexts, automation bias (cf. Section 4.2) is a likely-to-observe phenomenon where people tend to use the outcome of an automated decision entity "as a heuristic replacement for vigilant information seeking and processing" ([56] p. 205). This makes it particularly challenging for people to detect and adequately address system failures or erroneous outputs. Whereas research has identified several avenues to improve human abilities to oversee systems (e.g., training of oversight persons [2] or augmenting system outputs with confidence scores [2, 3]), the positive effects associated with these avenues seem to be inconsistent [31]. The criticism regarding human oversight could be expressed in an even more foundational way: The driving momentum behind many AI-based systems is the intention of producing decisions and actions that are, in some regard, better than what human experts could achieve [31, 89]. Ironically, the very human beings whose imperfections are meant to be overcome by such an AI-based system end up in the role of overseers of the system [89]. This is, for instance, a typical situation in certain areas of AI-based medical diagnostics [34, 58].

Positive effects of human oversight. However, it is also possible to draw more positive conclusions from the literature. There are relevant examples where human operators successfully identified errors made by automated systems [14, 54]. In addition, research has enhanced our understanding of factors that affect human abilities to detect errors made by automated systems [14, 46, 54, 71]. Furthermore, ongoing research contributes to successfully unveiling design options to improve human-system collaboration [46, 85]. For example, recent work proposes to combine the strengths of machine and human decision-making by using so-called *conformal predictions* for multi-label classifiers. There, the machine proposes a set of outputs (instead of a single output), which is guaranteed to contain the correct output with a probability of at least some pre-defined value [79]. If this set contains only a single output, confidence is considered high enough to let the decision happen fully autonomously. If, instead, it contains more than one output, the choice is delegated to a human. It was shown that for certain tasks, the human-system decision accuracy indeed is higher when the human choice is restricted to the outputs that are in the proposed set [78].

Overall, research reports mixed results regarding human oversight. However, the fact that effective human oversight is a popular requirement in legislation emphasizes the necessity of obtaining clarity on the very concept of effective human oversight. This clarity is crucial to better devise powerful laws and guidelines and for ensuring the successful and safe implementation of AI in high-risk contexts. Thus, in this paper, we aim to derive high-level conditions for effectiveness in human oversight.

Moral Responsibility. While the approach to effectiveness in human oversight that we suggest in this paper is the original result of collective insights of an interdisciplinary team of researchers from psychology, law, philosophy, and computer science, we also draw inspiration from several high-level conceptions of moral responsibility, as, e.g., in [26, 59, 82]. Most conceptions of moral responsibility have a focus on three aspects in common: a causal connection between the agent and the event for which the agent might be responsible, the agent having sufficient knowledge of their decision situation, and a certain autonomy of the agent [26, 59, 82]. This high-level conception is used as a rough guardrail in this paper.

3 WHAT IS EFFECTIVE HUMAN OVERSIGHT?

We understand human oversight to be the supervision of a system by at least one natural person, typically with the authority to influence its operations or effects.[2] This influence operates at various levels throughout the system's lifecycle, including intervention during execution, reversal of faulty decisions, or adjustments of parameters to improve results [55]. Although the oversight person may be assigned additional responsibilities, such as monitoring system performance or making engineering decisions, these tasks are not themselves part of human oversight. Even when having fixated a meaning for "human oversight", it is not clear what "effective" means in this regard.

3.1 An objective-first approach to the effectiveness of human oversight

We want oversight to be useful. Therefore, it makes sense to define the effectiveness of human oversight in terms of how well it achieves its objectives. As we have discussed above, risk mitigation can be taken to be one of the main objectives of human oversight. This is the objective we will be focusing on. Additionally, effective human oversight arguably may also yield other positive effects, such as improved incorporation of ethical considerations [19], or enhanced judgment in cases impacting individual well-being [7, 51].

It is important to note that the understanding of effectiveness we are about to suggest is not the only viable one. While offering

[2] While this understanding of human oversight will serve as our working definition, it is not the only possible candidate (cf. Section 6). For further analysis of the concept of human oversight, see also [49].

a structured framework, it is just one among potentially numerous conceptualizations. One way to change our conceptualization of effectiveness in human oversight would be to put objectives other than risk mitigation into focus, such as improved accuracy, trustworthiness, actual trust, human autonomy, accountability and responsibility, or liability. Discussing them is beyond the scope of this paper; however, *prima facie*, there appears to be no inherent conflict between our proposal and these objectives. Furthermore, we believe that our conceptualization will be helpful beyond risk mitigation since it will also be able to address other objectives of human oversight, including (moral) responsibility, as we will demonstrate in Section 3.2.2.

In line with the proposition that the main objective of effective human oversight is to mitigate risk, the main question is: *When does human oversight facilitate the mitigation of risks?* We propose this is the case if the oversight person meets the following four key conditions: causal power, epistemic access, self-control, and fitting intentions. We believe that this makes intuitive sense: Usually, risks can be reliably mitigated by a human only if there is an action available to them that averts the risk (*causal power*), they can actually perform that action (*self-control*), they know that the risk is imminent and which action could mitigate it (*epistemic access*), and they also want to mitigate it (*fitting intentions*). This intuition can be put more precisely, as is done in the remainder of this subsection. Additionally, Figure 1 displays the four conditions of effectiveness.

3.1.1 *Causal power.*

To be effective, the human must be able to make a relevant change. More precisely:

Definition 3.1 (Causal power). The agent has the power to establish a sufficient causal connection to the relevant aspect of the world.

Applied to the context of human oversight, this means that *the oversight person has the power to establish a causal connection to parts of the system or its effects, especially the risk.* For instance, this may include the ability to influence the system, including its components, operation, or outputs. Which kind of causal power is appropriate and necessary depends on the context at hand. For example, a factory robot assisting a human in assembly might need a stop button or a manual control mechanism. For an AI tool that assesses university applications and makes recommendations for the admission committee it might be enough if the oversight person can overwrite or disregard the system's output and then forward possible issues to the system provider.

In any case, it is not enough if the oversight person is a mere observer of the system. Without proper means to interfere with it or at least with its effects, they could only watch the manifestation of a risk without being able to mitigate it. Also, it is not sufficient if the human only has 'pseudo-options', i.e., if they can only *seemingly* influence the system but not really make a relevant difference.[3] Instances of such scenarios are malfunctioning stop buttons, or environments in which the human's actions are by default not considered.

3.1.2 *Epistemic access.*

To be effective, the oversight person must know what to do and when to do it. More precisely:

Definition 3.2 (Epistemic access). The agent has sufficient knowledge of their decision situation (also cf. [72]).

Applied to the context of human oversight, this means that *the oversight person has sufficient knowledge about the risk and how to mitigate it, especially knowing a risk mitigation action.* An oversight person will, for instance, need to have a sufficient understanding of what the system is doing, in which state it is, which possible risks and benefits there are, which means of influencing the system exist, which effects will result from possible interventions, and which of these effects are most desirable. In some cases, this might include knowledge of normative, social, or cultural domains, without which it would be infeasible for the oversight person to assess certain risks correctly. They do not need to have perfect knowledge, though. They might be unsure or even ignorant about some aspects of their decision situation – as long as they know enough to achieve their relevant objectives, they have enough epistemic access [76]. Notably, we assume that the oversight person has certain basic knowledge, such as what a stop button is or how to operate a computer on a consumer-level. One measure to improve the epistemic access of the oversight person could arguably be to employ suitable explainability and transparency measures [6, 9, 15, 40, 61, 70, 77].

3.1.3 *Self-control.*

To be effective, the oversight person must be in charge of their own doing. More precisely:

Definition 3.3 (Self-control). The agent can decide for any path of action in their decision situation and, if they do so, follow through with it.[4]

Applied to the context of human oversight, this means that *if there is a risk-mitigating path of action, the human can decide for that path, and if they do so, they actually perform the corresponding action(s).* For this, the oversight person needs to be in their right state of mind and free in their actions in a relevant sense. They also have to be in a position to retain attention and purposefully target their actions during the oversight. If, for example, the job of overseeing a fleet of autonomous vehicles is so uneventful and dull that a normal person cannot help but let their mind wander off from their task, the oversight would ultimately become ineffective. Only if oversight persons have self-control, they can purposefully address risks. Other reasons why an oversight person could lack self-control would be when they are severely sleep-deprived, under the influence of substances, or suffering from an active seizure.

An oversight person who lacks self-control might also simultaneously lack sufficient causal power or epistemic access (e.g., in case of a seizure). Nevertheless, self-control is distinct from the other two conditions, as an oversight person can have self-control without having causal power or epistemic access. In fact, self-control may, in many situations, be an antecedent of causal power and epistemic access – but not vice versa.

Sometimes, oversight persons will decide in split-seconds or will instinctively perform a certain move, and we typically regard this as self-controlled. For example, an oversight person of a factory robot who realizes that it is just milliseconds away from severely injuring a human worker might reflexively slam the stop button. We consider this to be a real decision, even if made in a split-second;

[3] Pseudo-options, though, are not to be confused with merely unattractive but real options, such as pushing a stop button for fun and thereby halting an entire production line for no reason.

[4] In the sense of *guidance control* in [26].

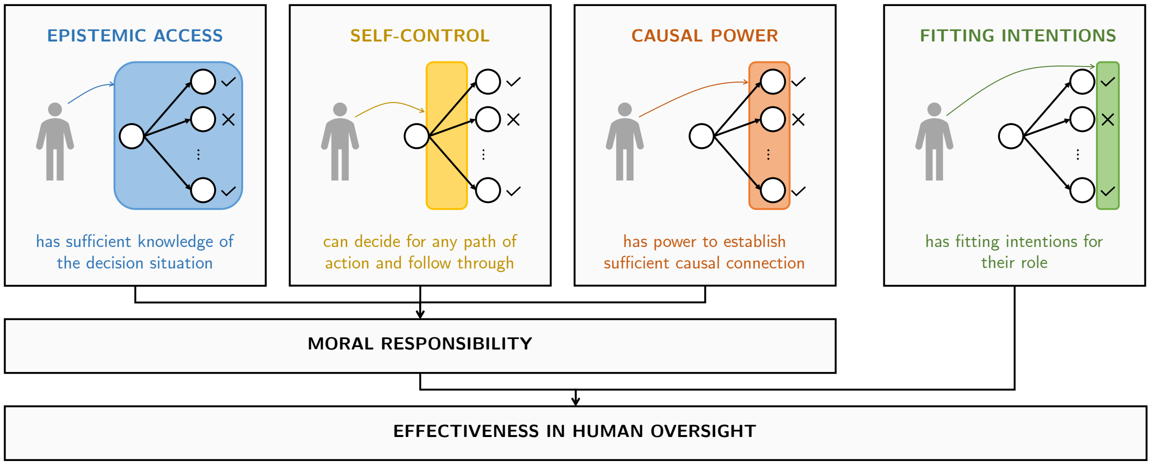
EPISTEMIC ACCESS
has sufficient knowledge of the decision situation
SELF-CONTROL
can decide for any path of action and follow through
CAUSAL POWER
has power to establish sufficient causal connection
FITTING INTENTIONS
has fitting intentions for their role
MORAL RESPONSIBILITY
EFFECTIVENESS IN HUMAN OVERSIGHT

Figure 1: The four conditions of effectiveness in human oversight and their relation to moral responsibility. The person in the diagram depicts the human in charge of oversight and the graph a schematic depiction of their current decision situation: the node on the left is their current position, the edges are the actions that they could perform, and the nodes on the right are the consequences of their actions.

and since the oversight person followed suit with their decision, they are to be considered self-controlled. Therefore, even oversight persons deciding instinctively can be effective according to the upcoming Definition 3.5.[5]

3.1.4 Fitting intentions. To be effective, the human must also want to do their job properly. More precisely:

Definition 3.4 (Fitting intentions). The agent has intentions that are fitting for their role (or some other relevant standard).

Applied to the context of human oversight, this means that *the oversight person* pro tanto *intends to mitigate risks, i.e., they intend to mitigate any risks while taking other relevant factors, such as the interests of the system's users and other stakeholders, into consideration in a suitable way.* If the oversight person were unwilling to make the efforts necessary to mitigate a risk, they would not be effective in their oversight – even if they had causal power, epistemic access, and self-control. This may, for example, be the case if the oversight person lacks the motivation to do their job properly or has other conflicting interests. Even worse, if an oversight person had ill intent, they could use their position to foster risks instead of mitigating them. For example, a racist overseeing a university admission system could deliberately overwrite outputs where people of color are ranked highly. In this case, their oversight should be considered ineffective in the relevant respects.

3.1.5 Effectiveness. The above yields the following conceptualization of effectiveness in human oversight:

Definition 3.5 (Effectiveness). An oversight person is effective in their human oversight if and only if they have *causal power*, *epistemic access*, *self-control*, and *fitting intentions* in the above senses.

Conceptualized this way, there still is considerable vagueness and generality to "effective human oversight", for example, due to the term "sufficient" in Definitions 3.1 and 3.2, or the "fitting" in Definition 3.4. This intentional vagueness is not a drawback but rather an essential feature. Not every oversight context requires the same level of epistemic access or causal power and, hence, the same exercise of oversight might count as effective in one context but not in another. Accordingly, our conceptualization of effectiveness allows for adaptability and versatility, acknowledging the complexity and diversity of situations where human oversight is applied.

Moreover, "effectiveness" is not a binary concept but a matter of degrees: one exercise of oversight might be *more effective* than another. Therefore, considering effectiveness in human oversight as a multifaceted continuum suitably encapsulates its nuanced nature. In line with the literature [31] and legislation [20], we nevertheless allow the use of "effective" as a binary concept. Saying that an oversight person is effective is to mean that they are *effective enough* for the concrete context they are operating in. In essence, this is also what Definition 3.5 says. While it is an open and important question of how much effectiveness is enough for a given context, we will not discuss this issue as it would go far beyond the scope of this paper.

3.2 Further considerations

3.2.1 Competence. Another aspect of an oversight person that could potentially come to mind is their *competence*. Intuitively, a competent oversight person will be better at being effective. While this is true, there is good reason why competence is not part of our definition of effectiveness: roughly speaking, someone is competent in performing a task if they know what to do to accomplish the task and are actually able to do it in a relevant sense. So, competence

[5] Alternatively, one might say there was not a true decision made in the example, but rather an instinctive jolt for the stop-button. In this case, the oversight person would not count as self-controlled according to Definition 3.3 (and therefore also not as effective according to Definition 3.5). If one wishes to allow this interpretation of the situation above and simultaneously allow that the human in it was exercising effective oversight, then the definition of self-control might be adapted to say "could" instead of "can". In this case, even certain instinctive movements of the above kind fall under the definition. However, since we want to avoid a discussion of the conceptual idealization introduced by the "could", we will stick to Definition 3.3 as is.

is tightly related to the conditions of effectiveness, even though competence is not itself part of the definition of effectiveness. A competent oversight person will be more likely to establish and retain sufficient epistemic access, self-control, and to use their causal powers in the right way. Without the necessary competence, an oversight person will arguably be unlikely to be effective, if they even can be effective at all.

3.2.2 Responsibility. Our conceptualization of effectiveness in human oversight already accounts for responsibility while also going beyond that. Taking inspiration from a definition of moral responsibility in computing in [59] and being consistent with the general idea of other sources [26, 82] we can define responsibility as follows: An agent is (morally) responsible for some aspect of the world if and only if (1) the agent has the power to establish a sufficient causal connection to this aspect of the world, (2) the agent has sufficient knowledge of their decision situation, and (3) the agent can decide for any path of action in their decision situation and, if they do so, follow through with it. In other words: an oversight person is morally responsible for risk mitigation (or any other part of their job) if and only if they have (1) causal power, (2) epistemic access, and (3) self-control in that respect. This can be easily translated to cover past events: an oversight person was responsible for, say, mitigating risk if and only if they had sufficient causal power, epistemic access, and self-control in that respect. These insights lead to the more concise formula that

An oversight person is effective if and only if they are morally responsible and have fitting intentions.[6]

This account also allows for shared responsibility between multiple agents, which is desirable in the context of high-risk AI systems. If, for example, both an oversight person and the deployer had causal power, epistemic access, and self-control with regard to a certain risk, they would both be responsible. For reasons of simplicity, the definition does not allow for group responsibility but can be adapted to do so.[7]

4 FACILITATORS AND INHIBITORS OF EFFECTIVE HUMAN OVERSIGHT

Beyond providing a general conceptualization, our proposed definition of effectiveness in human oversight has a practically useful dimension. Specifically, its conditions can stimulate ideas about possible facilitators and inhibitors of effectiveness. Facilitators are expected to promote effective oversight, while inhibitors are expected to hinder it. We organize them into three categories: technical design features, individual factors of oversight persons, and environmental circumstances (see, e.g., [63] for a similar scheme). The purpose of the discussion in this section is to provide readers with a more concrete understanding of possible factors regarding effective human oversight in accordance with our suggestions. The layout of Table 1 could serve as a template for thinking through possible facilitators and inhibitors, which can be useful to designers and deployers of AI-based systems. The following overview of facilitators and inhibitors is not meant to be exhaustive, and some of them may partially overlap each other. Also, they generally do not refer to a specific use case and some may not be applicable to every situation.

4.1 Technical design features

Under "technical design features", we subsume facilitators and inhibitors whose influence on effective human oversight results from the design of the system. This list could also include aspects such as interface and interaction design [60], runtime monitoring systems [8], fault tolerance [44], or automated anomaly detection systems [57].

Intervention options. For an oversight person to have causal power over the system, there must be options that allow the oversight person to intervene with, control, overwrite, or undo system decisions and actions. For example, an intervention option might be a stop button that allows oversight persons to stop system operation at any time. Other examples include options to take over manual control or to initiate emergency protocols that result in fail-safe modes.

System adaptability. It can be beneficial to oversight if oversight persons can adapt the system's processes or its interface [25, 69]. For example, the oversight person could reduce the speed at which the system produces its outputs or adapt the system's interface to display additional information. This increases the oversight person's epistemic access if the adaptations help him to acquire and process the information better.

System understandability. Under "system understandability" we subsume design aspects associated with the area of explainable AI [48]. For example, this includes the use of inherently transparent algorithms, the use of approaches to explain system outputs, the use of what-if analyses, or the provisioning of log files to enhance traceability of system processes [1, 75]. All options for enhancing system understandability aim at improving epistemic access [38, 48].

Interpretability of inputs and outputs. By "interpretability", we mean a representation of inputs or outputs that is appropriate for human understanding [67]. This could include, for example, the provision of information about what input features mean, or visualizing input and output features in a way that makes them easier to grasp. Interpretability thus aims at improving epistemic access.

Preselection of outputs to review. There may be cases where oversight tasks require a mechanism for selecting system outputs that are to be reviewed by the oversight person. By defining boundary conditions or by using oversight support tools, oversight persons could gain better epistemic access when only having to review a subset of outputs. For example, in a hiring context, such a tool could be used to identify applicants for whom the AI-based evaluation was obviously fair and free of bias, leaving considerably fewer cases to the oversight person [9].

[6] If Definition 3.3 were adapted to say "could", instead of "can" as suggested in footnote 5, this slogan would need to be altered to exclude cases of split-second, reflexive movements.

[7] Our conceptualization of responsibility should not be considered a blueprint for the legal domain. While it may share similarities with some (though not all) existing legal definitions of responsibility, caution is warranted against equating the two, since legal definitions of responsibility vary significantly across diverse areas of law, legal systems, and legal cultures.

	intervention options	system adaptability	system understandability	interpretability of in- and outputs	preselection of outputs to review	training of the oversight person	domain expertise	conscientiousness	exhaustion	motivation	automation bias	adequate job design	role conflicts	independent thinking	accountability	time pressure
	technical design					individual factors						environment				
causal power	•	•				•	•									○
epistemic access		•	•	•	•	•	•	•	○	•	○	•		•	•	○
self-control						•		•	○	•	○	•			•	
fitting intentions						•		•	○	•	○	•	○		•/○	

Table 1: Overview of how exemplary facilitators (•) and inhibitors (○) usually contribute to the effectiveness of human oversight.

4.2 Individual factors of oversight persons

Under "individual factors of oversight persons" we subsume traits and states of oversight persons as well as interventions *on* oversight persons that could facilitate or inhibit effective oversight. This discussion may cover aspects like vigilance [87], cognitive abilities [74, 88], humans' propensity to trust in automation [37], and additional cognitive biases [66] and heuristics [83].

Overseer Training. Training to prepare for or enhance the role of the oversight person will be crucial. For instance, training in which people are confronted with system errors or erroneous automated decisions can help them to become better at identifying these errors when later acting on the real system [2]. Overseer training could, in principle, target all of the conditions of effective human oversight. Such training could aim to enable oversight persons to maintain epistemic access, to recognize earlier when they are losing self-control (e.g., become inattentive [64]), to perform certain maneuvers that increase their causal power, or to know what the job requires them to do and thereby improve the fit of their intentions.

Domain expertise. Overseers may require a certain degree of expertise in the task performed by the system they are supposed to oversee. For instance, physicians with substantial expertise in diagnosing a disease may have better epistemic access (e.g., they will know better when a system is making incorrect diagnoses) than physicians with little expertise [29]. Domain expertise might sometimes also enable more causal power.

Conscientiousness. Conscientiousness is a character trait of people who are self-disciplined, orderly, goal-directed, who make and follow plans, and who adhere to social norms [12]. Highly conscientious oversight persons will arguably be more likely to maintain self-control, sufficient epistemic access, and fitting intentions.

Exhaustion. Exhaustion is one of many psychological strain symptoms [73]. It is a state where an oversight person feels tired and worn out [53] and thus may not be good at maintaining self-control and sufficient epistemic access.

Motivation. Motivation refers to internal and external factors that drive and direct an individual's behavior towards a goal [68]. Motivation leads people to initiate, intensify, and persist with goal-directed behavior. Motivation is often influenced by individual factors, environmental factors, and the perceived significance of the desired outcomes. In the context of human oversight, motivation is crucial to initiate and maintain fitting intentions and self-control for the goal of trying to mitigate risk. It is also important for persisting in maintaining sufficient epistemic access.

Automation bias. Automation bias is usually considered a cognitive bias where people use the outcome of an automated decision aid "as a heuristic replacement for vigilant information seeking and processing" ([56], p. 205). Automation bias becomes especially likely when facing a system that works reliably and with high performance because, in such cases, it is rational for people to divert their attention to other tasks [62]. This, however, will contribute to insufficient epistemic access. If there are indeed no negative consequences for some time, this may lead to what Parasuraman and Manzey [62] called "learned carelessness". This phenomenon could be interpreted as a problem of self-control or as an issue that leads to intentions that no longer fit the duties of an oversight person.

4.3 Environmental circumstances

Under "environmental circumstances", we subsume factors that lie outside the technical system or the oversight person, yet they can influence the effectiveness of oversight. This discussion could also include aspects such as stakes of the situation [47], single vs. multitasking environments [52], or whether oversight is organized as a team task [28, 84].

Adequate job design. The job of an oversight person needs to be well designed [63], which means that the job is, e.g., motivating, satisfying, and not overly demanding. For example, the job characteristics model from work psychology [33] highlights that certain job characteristics promote motivation and satisfaction with the job, namely skill variety, task identity, perceived task significance, autonomy, and feedback from the job.[8] The design of the job is

[8] See also other models from psychology that inspire job design to make the work more motivating and satisfying (e.g., [39]), or less demanding (e.g. the job-demands-resources model [16]).

important for promoting self-control, epistemic access, and fitting intentions, for example via motivation (cf. Section 4.2). Also, ill-designed jobs can lead to counterproductive work behavior, i.e., behavior at work that is inconsistent with or even intentionally disrespecting job duties (e.g., not even trying to maintain sufficient epistemic access, or intentionally overwriting outputs to feel "needed" [80]).

Role conflicts. Role conflicts could arise if the oversight person is at the same time a decision-maker who uses the AI-based system to support their decisions [43]. Consider, for example, a hiring scenario where an individual juggles both the role of an oversight person tasked with mitigating unfairness and the role of an HR manager who is responsible for inviting only the best candidates to a job interview. A conflict arises when the only method to address unfairness involves inviting additional, apparently less suitable candidates, conflicting with the HR manager's goal of selecting the best candidates only. This misalignment of duties may give rise to intentions that are unfitting for an oversight person.

Independent thinking. Processing a situation under the impression of a system's output can affect an oversight person's judgment of such an output [27]. For example, seeing a recidivism risk score for a defendant in court, in addition to the information about the defendant's case, may affect how people judge this case. Seeing a "high anchor" (i.e., the system's prediction of a high risk for recidivism) will make it more likely that the judge will decide that the defendant has a high risk for recidivism. Such anchors make oversight persons selectively aware of anchor-consistent information [18] which undermines epistemic access.

In the area of AI-supported decision-making, there are several ideas of how to possibly prevent such effects. For example, although this leads to less acceptance of the system, it seems to partly counter such anchoring effects if people are forced to think about a given case themselves for some time before making a final decision [11]. Alternatively, oversight persons may only receive the AI-based output after they have made up their own mind about a specific case because then they would not initially be affected by the AI-based output and may be in a better position to judge whether the AI-based output is appropriate [31, 89].

Accountability. Here we refer to accountability as "an obligation to explain and to justify [one's] conduct" [10, p.447]. If oversight persons feel accountable for their actions this may influence their attentiveness or may lead to them feeling more stressed [35]. Thus, a feeling of accountability could promote efforts of oversight persons to get themselves in a sufficient epistemic position. It can also be hypothesized that a felt obligation to explain oneself will motivate to maintain high self-control in order to keep up with the expectations of the job and that it will incentivize one to uphold intentions that fit one's duties. However, it could also lead to unintended consequences such as leading to unfit intentions: if oversight persons believe that they would have no good explanation for overriding a system output, they may not do so in cases where they know that they have to file a report after overwriting system outputs.

Time pressure. Significant time pressure may undermine various conditions for effectiveness [65]. For example, if an oversight person only has a few seconds to react to warnings by the system, this could hinder sufficient epistemic access. In extreme cases, time pressure could also undermine an oversight persons' causal power. For instance, if the option to overwrite an output is only available for split seconds, the oversight person may not be able to press the button.

5 APPLICATION TO THE AI ACT PROPOSAL

Finally, our analysis turns to effective human oversight as outlined in the AI Act [22][9]. According to Article 14(1) of the proposal, what the AI Act defines as high-risk AI systems shall be designed and developed in such a way that they can be effectively overseen during their use. As stated in Article 14(2), this oversight shall aim to prevent or minimize risks to health, safety, or fundamental rights. Risk mitigation is therefore defined as a key objective of human oversight measures that should contribute to the overall trustworthiness of high-risk systems [19]. Consequently, the effectiveness of human oversight, as outlined in Article 14(1), serves as a foundational principle that requires oversight measures to be capable of achieving their risk mitigation objective. This principle can be understood as a condition for a minimum level of effectiveness of the oversight measures implemented, comparable to a threshold that must be reached.

5.1 Conditions for effectiveness in our proposal and the AI Act

Article 14(3) obliges the provider to enable its deployer to perform effective oversight by identifying "appropriate measures".[10] Article 14(4) contains certain aspects of what the proposed AI Act envisages as indicators of the effectiveness of these measures. These require enabling the individuals to whom human oversight is assigned to do the following, as appropriate and proportionate to the circumstances:

> (a) to properly understand the relevant capacities and limitations of the high-risk AI system and be able to duly monitor its operation, including in view of detecting and addressing anomalies, dysfunctions and unexpected performance;
>
> (b) to remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
>
> (c) to correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available;
>
> (d) to decide, in any particular situation, not to use the high-risk AI system or to otherwise disregard, override or reverse the output of the high-risk AI system;

[9] At the time of writing, the trilogue negotiations between the European Commission, the Parliament, and the Council have been concluded and a finalized version has been approved by the European Parliament. However, the final version of the AI Act has not yet been approved by the Council of the EU.

[10] It should be noted that the Commission proposal [20] referred to the deployer as "user". The European Parliament [21] rightly proposed to use the term "deployer" instead to avoid any confusion with the end user of the AI system. This proposal found its way into the version approved by the Parliament [22]. We therefore use the term "deployer".

> (e) to intervene in the operation of the high-risk AI system or interrupt the system through a 'stop' button or a similar procedure that allows the system to come to a halt in a safe state.
>
> (quoted from Article 14(4) of the amendment to the version approved by the European Parliament [22])

These measures reflect our conditions of effectiveness. Requirements (d) and (e) underscore the EU's recognition of the importance of oversight persons being able to influence the AI system, either by modifying the system's output or by stopping its operation. These requirements manifest the need for both causal power and self-control, ensuring that individuals can influence the decisions of the AI system, or decide not to use the system's output or to modify it. Conditions (a) and (c) relate to the understanding of the system, as they require the provider to enable the oversight person to understand the limitations of the system (though not necessarily all details of the system), to monitor it effectively, and to interpret its results correctly. This is consistent with the condition of epistemic access, which enables the oversight person to make informed decisions. Awareness of automation bias, as described in (b), is intertwined with epistemic access and self-control, as described above in Section 4.2. Furthermore, Article 26(2) [22] emphasizes the need for the necessary competence, training, authority, and support in oversight functions, hence requiring epistemic access and causal power.[11] The criterion of fitting intentions of the agents involved has little direct equivalent in the legal text, which is not focused on subjective intentions but on actions. The parties' underlying intentions and motivations are not directly relevant. However, one might consider that the automation bias targeted in Article 14(4) (b) may interfere with the fittingness of the oversight person's intentions (cf. Section 4.2). Furthermore, the significant penalties according to Article 99 [22] strongly incentivize norm-compliant behavior.

5.2 Further analysis of the EU legislation's approach

Strongly varying levels of abstraction. It is striking that the measures appearing in Article 14(4) are of varying degrees of abstraction: While some measures contain specific requirements for the design of the AI system, other measures are less specific. For example, the requirement of a "stop" button in (e) is highly concrete. In contrast, the requirement in (a) to create an understanding of the limitations of the system is vague and leaves a lot of room for interpretation. Overall, the requirements for the various parties – providers of the AI system, deployers, and the person entrusted with oversight – could have been made clearer in the structure of the AI Act. Thus, Article 14(4) appears to be a loose collection of items that were deemed useful in the legislative process without much structural consistency. The lack of regulatory clarity could hinder the practical applicability of Article 14 and, thus, the introduction of truly effective human oversight.

Significance of Psychological Biases. Measures in accordance with (b), which are intended to maintain an awareness of the automation bias, seem out of place: it introduces a concrete psychological aspect that sticks out from the more technical and operational aspects outlined in the other provisions, especially considering that automation bias is just one of many relevant biases (e.g., hindsight bias, confidence bias [36, 66]), heuristics (e.g., availability heuristic [83]), and other psychological phenomena (e.g., exhaustion, motivation [68, 73]) that would need to be accounted for.

Legal uncertainties. Not every measure in Article 14(4) will have to be fully implemented in every high-risk system. This already follows from the restriction that the measures should be implemented "as appropriate and proportionate to the circumstances". As a result, the specific measures taken must be assessed on a case-by-case basis. Not only *how* these requirements apply, but also *which* requirements are applicable depends heavily on the individual case and the design of the system. Although AI Act's approach provides flexibility that accommodates the sometimes beneficial vagueness of the effectiveness of human oversight also mentioned in Section 3.1.5, it may also be associated with greater legal uncertainty. The legislation will eventually be complemented by technical standards developed by European standardization organizations to at least partially remove this legal uncertainty and to facilitate the implementation of requirements for high-risk systems, including human oversight [86]. However, this standardization process is still in its early stages and is limited to a few general guidelines [49]. Moreover, the question arises as to what requirements these standards must meet to satisfy the effectiveness requirement of Article 14(1). While Article 14(4) offers some guidance, its vagueness and varying levels of abstraction are insufficient for determining true effectiveness in specific cases. Thereby, the standardization organisations risk developing technical standards that may not withstand legal scrutiny. So, the EU's reliance on the AI Act's effective implementation through practical technical standards poses a potential threat to the Regulation's requirement for human oversight.

The practical need for a definition. Also, the technical standards created in this process only have a presumptive effect for providers and deployers of AI systems that are subject to the AI Act. In the event of a legal dispute, the courts will therefore have to examine whether the specific design of the human oversight measures meets the overarching requirement of effectiveness in Article 14(1). As mentioned above, this will depend on the circumstances of the individual case. To ensure uniform application of the AI Act by lower courts, a comprehensible definition of effective human oversight is essential. In this respect, the definition of effectiveness will be one of the key tasks of the deciding courts in the coming years, leaving these courts also in need of further clarification on the conditions of effectiveness.

5.3 Utilizing our definition of effectiveness

Our objective-first approach to effectiveness can be used to integrate the piecemeal list of different requirements listed in the AI Act into a coherent system. This is advantageous for a legally sound application of the AI Act, as it helps to structure the requirements for and responsibilities of the different parties involved. One of

[11] This requirement was mentioned only in the non-binding recital (48) of the EU Commission's proposal [20], but was ultimately stipulated more directly in Article 29 of the proposals by the EU Council [13], the Parliament [21] and Article 26 of the approved version [23].

the benefits of the breadth and structure of our conceptualization is that it simplifies the assessment process for regulators, making it easier to evaluate different AI systems against a structured set of principles that are framed by the four conditions. The three domains of facilitators and inhibitors can aid in giving further structure for certain use cases. This streamlining may also decrease the legal uncertainty in Article 14, fostering a more consistent and predictable environment for AI providers and deployers alike, while providing a practical definition of effectiveness for deciding courts. This could also help uncover further shortcomings in Article 14(4): For instance, while understanding the system's output (c) can enhance epistemic access, understanding of the system's *input* can be equally crucial. Our conditions for effectiveness, when applied with expertise, make this need evident; whereas this need can be challenging to identify when relying solely on Article 14(4), which does not mention the system's input at all.

Consequently, the points listed in Table 1 can be understood as examples of how an assessment of the effectiveness of measures could be carried out. Defining effectiveness using our four conditions provides a practical framework that highlights the key aspects of effective oversight. Our approach is, therefore, well suited to assist standard-setting organizations and to clarify the rather broad requirements of Article 14(4). In addition, our proposed framework provides guidance to practitioners who do not wish to rely solely on technical standards. Finally, our approach can assist courts in the future by providing a framework for defining and assessing the effectiveness of human oversight.

6 LIMITATIONS AND FUTURE WORK

As discussed above, we see advantages brought about by our conceptualization of effective human oversight. Nevertheless, we also want to point out three limitations as well as future points to address. *First*, our conceptualization is designed for broad applicability across various AI-based systems, and thus shares a common criticism with the AI Act, namely that of residing on a high level of abstraction and that of vagueness, as discussed in Section 3.1.5. Dealing with these issues, especially in complex scenarios, requires (a) involving individuals with extensive expertise in the task at hand and (b) fostering collaboration in multidisciplinary teams to assess the impact on human oversight practice. *Second*, when we came up with the facilitators and inhibitors spelled out in Section 4, we realized that our conceptualization is helpful for thinking through conditions that could promote or hinder effective oversight. However, for some of the facilitators (or inhibitors), the question of which of the four conditions they contribute to (or hinder) depends on the concrete context. Moreover, for certain cases, the mapping from facilitators and inhibitors to conditions needs considerable empirical research, outside the scope of this conceptual paper. *Third*, we did not discuss the concept of human oversight itself in much detail, but focused only *effectiveness* of human oversight. So, while our paper is helpful in determining whether an instance of human oversight is effective, it is less helpful in assessing whether something is human oversight. However, our definition of effectiveness is broad enough to be applied to many roles along the AI pipeline, not only to oversight persons. Some more conceptual discussion of human oversight itself is, for instance, provided by [49].

Future efforts will aim to enhance the practical applicability of our framework. We plan to develop a comprehensive practical framework, help to establish best practices, and create tools for implementing and auditing effective human oversight. These developments can benefit both policymakers and practitioners. This endeavor will require ongoing interdisciplinary collaboration, such as conducting empirical studies to validate and refine oversight mechanisms and designing holistic requirements for high-risk systems, oversight personnel, and their environments. Additionally, efforts should include creating educational programs and training for oversight persons, as well as engaging with regulatory bodies to ensure the ethical deployment of AI systems.

7 CONCLUSION

This paper has proposed an approach to effective human oversight – encompassing causal power, epistemic access, self-control, and fitting intentions – to remedy the lack of clarity surrounding the concept of effectiveness. This is a joint contribution of researchers in psychology, law, philosophy, and computer science. We argued that, in essence, a morally responsible oversight person with fitting intentions is generally suitable for mitigating risks associated with high-risk AI systems. We have identified facilitators and inhibitors of effectiveness in three categories, namely the technical design of the system, individual characteristics of oversight persons, and the environmental circumstances in which oversight occurs. Thereby, we have provided inspiration for steps towards a successful implementation of effective human oversight. We also discussed the extent to which our understanding of effective human oversight aligns with the European Union's upcoming AI Act, in particular Article 14 on human oversight. Our analysis suggests that the AI Act could benefit from incorporating more nuanced and structured conditions, such as those proposed in this paper.

In closing, we want to advocate for a balanced approach to effective human oversight. Without a concrete idea of what effectiveness is in the context of human oversight, no productive discussion of it is possible. On the one hand, we urge to not throw out the baby with the bathwater by per se dismissing human oversight as a safeguard without the necessary conceptual clarity. On the other hand, caution is warranted as well when emphatically including a notion of effective human oversight in regulations and guidelines without this conceptual clarity. Our interdisciplinary approach is intended to aid both advocates and critics of human oversight. For many relevant oversight contexts, the four conditions of effectiveness seem to provide a better picture of when and how human oversight is useful. We, therefore, believe that these conditions could form a pivot for future research on human oversight.

ACKNOWLEDGMENTS

This work is partially funded by DFG grant 389792660 as part of TRR 248 – CPEC, by VolkswagenStiftung as part of grants AZ 98514, 98513 and 98512 – EIS, and by the European Regional Development Fund and the Saarland within the scope of (To)CERTAIN.

REFERENCES

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI):

Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[2] J. Elin Bahner, Monika F. Elepfandt, and Dietrich Manzey. 2008. Misuse of Diagnostic Aids in Process Control: The Effects of Automation Misses on Complacency and Automation Bias: (578262012-006). https://doi.org/10.1037/e578262012-006

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (Oct. 2019), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM. https://doi.org/10.1145/3411764.3445717

[5] Megan L. Bartlett and Jason S. McCarley. 2020. Ironic efficiency in automation-aided signal detection. *Ergonomics* 64, 1 (Aug. 2020), 103–112. https://doi.org/10.1080/00140139.2020.1809716

[6] Deborah Baum, Kevin Baum, Timo P. Gros, and Verena Wolf. 2023. XAI Requirements in Smart Production Processes: A Case Study. In *Explainable Artificial Intelligence*, Luca Longo (Ed.). Springer Nature Switzerland, Cham, 3–24.

[7] Kevin Baum, Susanne Mantel, Eva Schmidt, and Timo Speith. 2022. From Responsibility to Reason-Giving Explainable Artificial Intelligence. *Philosophy & Technology* 35, 1 (2022), 12. https://doi.org/10.1007/s13347-022-00510-w

[8] Jan Baumeister, Bernd Finkbeiner, Sebastian Schirmer, Maximilian Schwenger, and Christoph Torens. 2020. RTLola Cleared for Take-Off: Monitoring Autonomous Aircraft. In *CAV 2020 (LNCS, Vol. 12225)*. Springer, 28–39. https://doi.org/10.1007/978-3-030-53291-8_3

[9] Sebastian Biewer, Kevin Baum, Sarah Sterz, Holger Hermanns, Sven Hetmank, Markus Langer, Anne Lauber-Rönsberg, and Franz Lehr. 2024. Software doping analysis for human oversight. *Formal Methods in System Design* (April 2024). https://doi.org/10.1007/s10703-024-00445-2

[10] Mark Bovens. 2007. Analysing and Assessing Accountability: A Conceptual Framework1. *European Law Journal* 13, 4 (June 2007), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

[11] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (April 2021), 1–21. https://doi.org/10.1145/3449287

[12] Paul T. Costa, Robert R. McCrae, and David A. Dye. 1991. Facet Scales for Agreeableness and Conscientiousness: A Revision of the NEO Personality Inventory. *Personality and Individual Differences* 12, 9 (1991), 887–898. https://doi.org/10.1016/0191-8869(91)90177-d

[13] Council of the European Union. 2022. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts – General Approach, 14954/22. https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf.

[14] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM. https://doi.org/10.1145/3313831.3376638

[15] Luca Deck, Astrid Schoemäcker, Timo Speith, Jakob Schöffer, Lena Kästner, and Niklas Kühl. 2024. Mapping the Potential of Explainable Artificial Intelligence (XAI) for Fairness Along the AI Lifecycle. *arXiv preprint arXiv:2404.18736* (2024).

[16] Evangelia Demerouti, Arnold B. Bakker, Friedhelm Nachreiner, and Wilmar B. Schaufeli. 2001. The job demands-resources model of burnout. *Journal of Applied Psychology* 86, 3 (2001), 499–512. https://doi.org/10.1037//0021-9010.86.3.499

[17] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22)*. ACM. https://doi.org/10.1145/3491101.3503727

[18] Birte Englich, Thomas Mussweiler, and Fritz Strack. 2006. Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. *Personality and Social Psychology Bulletin* 32, 2 (Feb. 2006), 188–200. https://doi.org/10.1177/0146167205282152

[19] Lena Enqvist. 2023. 'Human oversight'in the EU artificial intelligence act: what, when and by whom? *Law, Innovation and Technology* 15, 2 (2023), 508–535.

[20] European Commission. 2021. Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (proposal for a regulation) no 0106/2021. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[21] European Parliament. 2023. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), P9_TA(2023)0236. https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf.

[22] European Parliament. 2024. Corrigendum to the position of the European Parliament adopted at first reading on 13 March 2024 with a view to the adoption of Regulation (EU) 2024/ ...... of the European Parliament and of the Council laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) P9_TA(2024)0138 (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf.

[23] European Parliament. 2024. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)), P9_TA(2024)0138. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

[24] European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679.

[25] João Marcelo Evangelista Belo, Mathias N. Lystbæk, Anna Maria Feit, Ken Pfeuffer, Peter Kán, Antti Oulasvirta, and Kaj Grønbæk. 2022. AUIT – the Adaptive User Interfaces Toolkit for Designing XR Applications. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. ACM. https://doi.org/10.1145/3526113.3545651

[26] John Martin Fischer and Mark Ravizza. 1998. *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge University Press, New York.

[27] Adrian Furnham and Hua Chu Boo. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics* 40, 1 (Feb. 2011), 35–42. https://doi.org/10.1016/j.socec.2010.10.008

[28] Fei Gao, Mary L. Cummings, and Erin Treacy Solovey. 2014. Modeling Teamwork in Supervisory Control of Multiple Robots. *IEEE Transactions on Human-Machine Systems* 44, 4 (Aug. 2014), 441–453. https://doi.org/10.1109/thms.2014.2312391

[29] Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo K. Koch, Matthias F. C. Hudecek, Alun D. Ackery, Samir C. Grover, Joseph F. Coughlin, Dieter Frey, Felipe C. Kitamura, Marzyeh Ghassemi, and Errol Colak. 2023. Non-task expert physicians benefit from correct explainable AI advice when reviewing X-rays. *Scientific Reports* 13, 1 (Jan. 2023). https://doi.org/10.1038/s41598-023-28633-w

[30] Government of Canada. 2023. Directive on Automated Decision-Making. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32746&section=html

[31] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45 (2022), 105681. https://doi.org/10.1016/j.clsr.2022.105681

[32] Ben Green and Yiling Chen. 2019. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM. https://doi.org/10.1145/3287560.3287563

[33] J.Richard Hackman and Greg R. Oldham. 1976. Motivation through the design of work: test of a theory. *Organizational Behavior and Human Performance* 16, 2 (Aug. 1976), 250–279. https://doi.org/10.1016/0030-5073(76)90016-7

[34] Sarah Haggenmüller, Roman C. Maron, Achim Hekler, Jochen S. Utikal, Catarina Barata, Raymond L. Barnhill, Helmut Beltraminelli, Carola Berking, Brigid Betz-Stablein, Andreas Blum, Stephan A. Braun, Richard Carr, Marc Combalia, Maria-Teresa Fernandez-Figueras, Gerardo Ferrara, Sylvie Fraitag, Lars E. French, Frank F. Gellrich, Kamran Ghoreschi, Matthias Goebeler, Pascale Guitera, Holger A. Haenssle, Sebastian Haferkamp, Lucie Heinzerling, Markus V. Heppt, Franz J. Hilke, Sarah Hobelsberger, Dieter Krahl, Heinz Kutzner, Aimilios Lallas, Konstantinos Liopyris, Mar Llamas-Velasco, Josep Malvehy, Friedegund Meier, Cornelia S.L. Müller, Alexander A. Navarini, Cristián Navarrete-Dechent, Antonio Perasole, Gabriela Poch, Sebastian Podlipnik, Luis Requena, Veronica M. Rotemberg, Andrea Saggini, Omar P. Sangueza, Carlos Santonja, Dirk Schadendorf, Bastian Schilling, Max Schlaak, Justin G. Schlager, Mildred Sergon, Wiebke Sondermann, H. Peter Soyer, Hans Starz, Wilhelm Stolz, Esmeralda Vale, Wolfgang Weyers, Alexander Zink, Eva Krieghoff-Henning, Jakob N. Kather, Christof von Kalle, Daniel B. Lipka, Stefan Fröhling, Axel Hauschild, Harald Kittler, and Titus J. Brinker. 2021. Skin cancer classification via convolutional neural networks: systematic review of studies involving human experts. *European Journal of Cancer* 156 (2021), 202–216. https://doi.org/10.1016/j.ejca.2021.06.049

[35] Angela T. Hall, Dwight D. Frink, and M. Ronald Buckley. 2015. An accountability account: A review and synthesis of the theoretical and empirical research on felt accountability. *Journal of Organizational Behavior* 38, 2 (Sept. 2015), 204–224. https://doi.org/10.1002/job.2052

[36] Martin Hilbert. 2012. Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin* 138, 2 (March 2012), 211–237. https://doi.org/10.1037/a0025940

[37] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[38] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[39] Stephen E. Humphrey, Jennifer D. Nahrgang, and Frederick P. Morgeson. 2007. Integrating motivational, social, and contextual work design features: A meta-analytic summary and theoretical extension of the work design literature. *Journal of Applied Psychology* 92, 5 (2007), 1332–1356. https://doi.org/10.1037/0021-9010.92.5.1332

[40] Jinglu Jiang, Surinder Kahai, and Ming Yang. 2022. Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies* 165 (2022), 102839. https://doi.org/10.1016/j.ijhcs.2022.102839

[41] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (Sept. 2019), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[42] Meg Leta Jones. 2017. The right to a human in the loop: Political constructions of computer automation and personhood. *Social Studies of Science* 47, 2 (2017), 216–239. https://doi.org/10.1177/0306312717699716 PMID: 28406392.

[43] Lynda A. King and Daniel W. King. 1990. Role conflict and role ambiguity: A critical assessment of construct validity. *Psychological Bulletin* 107, 1 (1990), 48–64. https://doi.org/10.1037//0033-2909.107.1.48

[44] Israel Koren and C Mani Krishna. 2020. *Fault-tolerant systems*. Morgan Kaufmann.

[45] Riikka Koulu. 2020. Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastricht Journal of European and Comparative Law* 27 (2020), 720 – 735. https://api.semanticscholar.org/CorpusID:229412494

[46] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM. https://doi.org/10.1145/3593013.3594087

[47] Markus Langer and Richard N. Landers. 2021. The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior* 123 (Oct. 2021), 106878. https://doi.org/10.1016/j.chb.2021.106878

[48] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (July 2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[49] Johann Laux. 2023. Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. *AI & SOCIETY* (Oct. 2023). https://doi.org/10.1007/s00146-023-01777-z

[50] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[51] Chiara Longoni, Andrea Bonezzi, and Carey K Morewedge. 2019. Resistance to Medical Artificial Intelligence. *Journal of Consumer Research* 46, 4 (May 2019), 629–650. https://doi.org/10.1093/jcr/ucz013

[52] Joseph B. Lyons and Svyatoslav Y. Guznov. 2018. Individual differences in human–machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science* 20, 4 (Nov. 2018), 440–458. https://doi.org/10.1080/1463922x.2018.1491071

[53] Christina Maslach, Susan E Jackson, and Michael P Leiter. 1997. *Maslach burnout inventory*. Scarecrow Education.

[54] Sara E. McBride, Wendy A. Rogers, and Arthur D. Fisk. 2013. Understanding human management of automation errors. *Theoretical Issues in Ergonomics Science* 15, 6 (Aug. 2013), 545–577. https://doi.org/10.1080/1463922x.2013.817625

[55] Leila Methnani, Andrea Aler Tubella, Virginia Dignum, and Andreas Theodorou. 2021. Let Me Take Over: Variable Autonomy for Meaningful Human Control. *Frontiers in Artificial Intelligence* 4 (2021). https://doi.org/10.3389/frai.2021.737072

[56] Kathleen L. Mosier, Linda J. Skitka, Mark D. Burdick, and Susan T. Heers. 1996. Automation Bias, Accountability, and Verification Behaviors. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40, 4 (Oct. 1996), 204–208. https://doi.org/10.1177/154193129604000413

[57] Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, and Fatima Mohamad Dakalbab. 2021. Machine Learning for Anomaly Detection: A Systematic Review. *IEEE Access* 9 (2021), 78658–78700. https://doi.org/10.1109/ACCESS.2021.3083060

[58] Naoshi Nishida, Makoto Yamakawa, Tsuyoshi Shiina, Yoshito Mekada, Mutsumi Nishida, Naoya Sakamoto, Takashi Nishimura, Hiroko Iijima, Toshiko Hirai, Ken Takahashi, Masaya Sato, Ryosuke Tateishi, Masahiro Ogawa, Hideaki Mori, Masayuki Kitano, Hidenori Toyoda, Chikara Ogawa, and Masatoshi Kudo. 2022. Artificial intelligence (AI) models for the ultrasonographic diagnosis of liver tumors and comparison of diagnostic accuracies between AI and human experts. *Journal of Gastroenterology* 57, 4 (Feb. 2022), 309–321. https://doi.org/10.1007/s00535-022-01849-9

[59] Merel Noorman. 2020. Computing and Moral Responsibility. In *The Stanford Encyclopedia of Philosophy* (Spring 2020 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[60] Donald A. Norman. 1983. Design principles for human-computer interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 1983, Boston, Massachusetts, USA, December 12-15, 1983*, Raoul N. Smith, Richard W. Pew, and Ann Janda (Eds.). ACM, 1–10. https://doi.org/10.1145/800045.801571

[61] Daria Onitiu. 2023. The limits of explainability & human oversight in the EU Commission's proposal for the Regulation on AI- a critical approach focusing on medical diagnostic systems. *Information & Communications Technology Law* 32, 2 (2023), 170–188. https://doi.org/10.1080/13600834.2022.2116354 arXiv:https://doi.org/10.1080/13600834.2022.2116354

[62] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52, 3 (June 2010), 381–410. https://doi.org/10.1177/0018720810376055

[63] Sharon K. Parker and Gudela Grote. 2020. Automation, Algorithms, and Beyond: Why Work Design Matters More Than Ever in a Digital World. *Applied Psychology* 71, 4 (Feb. 2020), 1171–1204. https://doi.org/10.1111/apps.12241

[64] Franziska Perels, Tina Gürtler, and Bernhard Schmitz. 2005. Training of self-regulatory and problem-solving competence. *Learning and Instruction* 15, 2 (April 2005), 123–139. https://doi.org/10.1016/j.learninstruc.2005.04.010

[65] Tobias Rieger and Dietrich Manzey. 2022. Understanding the Impact of Time Pressure and Automation Support in a Visual Search Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society* (June 2022), 001872082211112. https://doi.org/10.1177/00187208221111236

[66] Neal J. Roese and Kathleen D. Vohs. 2012. Hindsight Bias. *Perspectives on Psychological Science* 7, 5 (Sept. 2012), 411–426. https://doi.org/10.1177/1745691612454303

[67] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (May 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[68] Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology* 25, 1 (Jan. 2000), 54–67. https://doi.org/10.1006/ceps.1999.1020

[69] Juergen Sauer and Alain Chavaillaz. 2017. How operators make use of wide-choice adaptable automation: observations from a series of experimental studies. *Theoretical Issues in Ergonomics Science* 19, 2 (March 2017), 135–155. https://doi.org/10.1080/1463922x.2017.1297866

[70] Nadine Schlicker, Markus Langer, Sonja K. Ötting, Kevin Baum, Cornelius J. König, and Dieter Wallach. 2021. What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. *Comput. Hum. Behav.* 122 (2021), 106837. https://doi.org/10.1016/j.chb.2021.106837

[71] Jakob Schoeffer, Johannes Jakubik, Michael Voessing, Niklas Kuehl, and Gerhard Satzger. 2023. *On the Interdependence of Reliance Behavior and Accuracy in AI-Assisted Decision-Making*. IOS Press. https://doi.org/10.3233/faia230074

[72] Michael J. Shaffer. 2012. *Epistemic Access, Confirmation, and Idealization*. Palgrave Macmillan UK, London, 101–144. https://doi.org/10.1057/9781137271587_4

[73] Sabine Sonnentag, Louis Tay, and Hadar Nesher Shoshan. 2023. A review on health and well-being at work: More than stressors and strains. *Personnel Psychology* 76, 2 (Jan. 2023), 473–510. https://doi.org/10.1111/peps.12572

[74] Charles Spearman. 1904. " General Intelligence" Objectively Determined and Measured. 15 (1904), 201–292.

[75] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM. https://doi.org/10.1145/3531146.3534639

[76] T. Speith, B. Crook, S. Mann, A. Schomäcker, and M. Langer. In press. Conceptualizing Understanding in Explainable Artificial Intelligence (XAI): An Abilities-Based Approach. *Ethics and Information Technology* (In press).

[77] Sarah Sterz, Kevin Baum, Anne Lauber-Rönsberg, and Holger Hermanns. 2021. Towards Perspicuity Requirements. In *29th IEEE International Requirements Engineering Conference Workshops, RE 2021 Workshops, Notre Dame, IN, USA, September 20-24, 2021*, Tao Yue and Mehdi Mirakhorli (Eds.). IEEE, 159–163. https://doi.org/10.1109/REW53955.2021.00029

[78] Eleni Straitouri and Manuel Gomez Rodriguez. 2023. Designing Decision Support Systems Using Counterfactual Prediction Sets. *CoRR* abs/2306.03928 (2023). https://doi.org/10.48550/ARXIV.2306.03928 arXiv:2306.03928

[79] Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. 2023. Improving Expert Predictions with Conformal Prediction. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 32633–32653. https://proceedings.mlr.press/v202/straitouri23a.html

[80] Franz Strich, Anne-Sophie Mayer, and Marina Fiedler. 2021. What Do I Do in a World of Artificial Intelligence? Investigating the Impact of Substitutive Decision-Making AI Systems on Employees' Professional Role Identity. *Journal of the Association for Information Systems* 22, 2 (2021), 304–324. https://doi.org/10.17705/1jais.00663

[81] Mohammad Tahaei, Marios Constantinides, Daniele Quercia, Sean Kennedy, Michael Muller, Simone Stumpf, Q Vera Liao, Ricardo Baeza-Yates, Lora Aroyo,

Jess Holbrook, et al. 2023. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–4.

[82] Matthew Talbert. 2019. Moral Responsibility. In *The Stanford Encyclopedia of Philosophy* (Winter 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[83] Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5, 2 (Sept. 1973), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

[84] Anna-Sophie Ulfert, Eleni Georganta, Carolina Centeio Jorge, Siddharth Mehrotra, and Myrthe Tielman. 2023. Shaping a multidisciplinary understanding of team trust in human-AI teams: a theoretical framework. *European Journal of Work and Organizational Psychology* (April 2023), 1–14. https://doi.org/10.1080/1359432x.2023.2200172

[85] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–38. https://doi.org/10.1145/3579605

[86] Michael Veale and Frederik J. Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act. *CoRR* abs/2107.03721 (2021). arXiv:2107.03721 https://arxiv.org/abs/2107.03721

[87] Joel S. Warm, Raja Parasuraman, and Gerald Matthews. 2008. Vigilance Requires Hard Mental Work and Is Stressful. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 3 (June 2008), 433–441. https://doi.org/10.1518/001872008x312152

[88] Russell T. Warne and Cassidy Burningham. 2018. Spearman's g Found in 31 Non-Western Nations: Strong Evidence that g is a Universal Phenomenon. (March 2018). https://doi.org/10.31234/osf.io/uv673

[89] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. 2019. Algorithmic Decision-Making and the Control Problem. *Minds and Machines* 29, 4 (Dec. 2019), 555–578. https://doi.org/10.1007/s11023-019-09513-7

[90] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. ACM. https://doi.org/10.1145/3351095.3372852