Design Considerations for Human Oversight of Al: Insights from Co-Design Workshops and Work Design Theory

CEDRIC FAAS, Saarland Informatics Campus, Saarland University, Germany SOPHIE KERSTAN, Department of Psychology, University of Freiburg, Germany RICHARD UTH, Department of Psychology, University of Freiburg, Germany MARKUS LANGER, Department of Psychology, University of Freiburg, Germany ANNA MARIA FEIT, Saarland Informatics Campus, Saarland University, Germany

As AI systems become increasingly capable and autonomous, domain experts' roles are shifting from performing tasks themselves to overseeing AI-generated outputs. Such oversight is critical, as undetected errors can have serious consequences or undermine the benefits of AI. Effective oversight, however, depends not only on detecting and correcting AI errors but also on the motivation and engagement of the oversight personnel and the meaningfulness they see in their work. Yet little is known about how domain experts approach and experience the oversight task and what should be considered to design effective and motivational interfaces that support human oversight. To address these questions, we conducted four co-design workshops with domain experts from psychology and computer science. We asked them to first oversee an AI-based grading system, and then discuss their experiences and needs during oversight. Finally, they collaboratively prototyped interfaces that could support them in their oversight task. Our thematic analysis revealed four key user requirements: understanding tasks and responsibilities, gaining insight into the AI's decision-making, contributing meaningfully to the process, and collaborating with peers and the AI. We integrated these empirical insights with the SMART model of work design to develop a generalizable framework of twelve design considerations. Our framework links interface characteristics and user requirements to the psychological processes underlying effective and satisfying work. Being grounded in work design theory, we expect these considerations to be applicable across domains and discuss how they extend existing guidelines for human-AI interaction and theoretical frameworks for effective human oversight by providing concrete guidance on the design of engaging and meaningful interfaces that support human oversight of AI systems.

CCS Concepts: • **Do Not Use This Code** → **Generate the Correct Terms for Your Paper**; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Additional Key Words and Phrases: Do, Not, Us, This, Code, Put, the, Correct, Terms, for, Your, Paper

1 Introduction

AI systems are increasingly capable of autonomous action and decision-making and are more and more deployed to enhance performance, improve quality, or enable new types of applications [see 43, for an overview]. However, in high-risk contexts such as medicine or in the educational or judicial system, there are practical, ethical, and legal reasons for adding one or multiple layers of human oversight to automated AI operations. Practically, humans can serve as a fallback when AI systems malfunction or fail and can detect inaccurate or inadequate (e.g., discriminatory) outputs to prevent risks to safety, health, or fundamental human rights. Ethically, human oversight ensures accountability and preserves human judgment in decisions affecting people's lives [21, 40]. Finally, legal frameworks, such as the European AI Act, mandate human oversight for high-risk applications that affect safety, health, or fundamental rights.

Authors' Contact Information: Cedric Faas, faas@cs.uni-saarland.de, Saarland Informatics Campus, Saarland University, Saarbrücken, Germany; Sophie Kerstan, sophie.kerstan@psychologie.uni-freiburg.de, Department of Psychology, University of Freiburg, Freiburg im Breisgau, Germany; Richard Uth, richard.bergs@psychologie.uni-freiburg.de, Department of Psychology, University of Freiburg, Freiburg im Breisgau, Germany; Markus Langer, markus.langer@psychologie.uni-freiburg.de, Department of Psychology, University of Freiburg, Freiburg im Breisgau, Germany; Anna Maria Feit, feit@cs.uni-saarland.de, Saarland Informatics Campus, Saarland University, Saarbrücken, Germany.

If human oversight is required, the conditions for it must be designed to make oversight *effective*. Effective oversight means that human overseers are able to detect inaccurate or inadequate system outputs and malfunctions and intervene appropriately to prevent possible negative effects [45]. This effectiveness is determined by the sociotechnical conditions of human oversight [75], which crucially includes the user interface that human oversight personnel use to monitor and intervene in AI operations. Prior work in explainability [2, 46, 52], and human-AI interaction design [33, 85], but also research on traditional automated systems and insights from the human factors community [22, 28, 44, 62, 70], offer valuable insights into how such interfaces need to be designed to enable understanding, detection, and control for humans to be effective in their oversight tasks.

However, oversight is not only a question of technical effectiveness but also of what organizational psychology calls work design, that is "the content and organization of one's work tasks, activities, relationships, and responsibilities" [61, p. 662] and how they affect workers' psychological needs, well-being, and job satisfaction. Oversight personnel must monitor AI systems for extended periods, which can make their roles monotonous, stressful, or detached from the core of their expertise, in particular, as human oversight of AI systems might substantially shift the responsibilities of domain experts, requiring them to transition from performing tasks based on their own judgment to overseeing AI systems and intervening effectively. Interfaces must therefore not only support accurate monitoring and timely intervention but also foster engagement, meaning, and a sense of agency [75]. Designing for human oversight therefore requires designing for "good work", those responsible for oversight stay engaged and motivated in their roles..

In this paper, we thus adopt a human-centered perspective on human oversight and ask: Which aspects should we consider when designing oversight interfaces to help domain experts be effective, motivated and engaged in overseeing AI systems? To address this question, we combined a participatory design approach with theoretical integration from work design. We first conducted four co-design workshops in which domain experts from different backgrounds (i.e., computer science, psychology) oversaw an AI-based grading system for student tests. Reflecting on their own approaches and experiences during the oversight task, participants discussed their needs and requirements for an oversight interface. They then critically engaged with these requirements by designing concrete paper prototypes.

Our qualitative thematic analysis revealed that participants often approached the oversight task by checking every individual test to assess AI, an approach they found ineffective and frustrating. Approaches that participants perceived as more positive included detecting unjustly awarded points by understanding the AI's strengths and weaknesses, ensuring fair grading of borderline cases, and focusing on tests that were easy and enjoyable to grade. Reflection on their approaches revealed four key user requirements: (1) having a clear understanding of their role, responsibilities, and tasks, (2) obtaining insights on the AI's capabilities and decision process, (3) contributing meaningfully to the task process, and (4) exchanging with other oversight personnel and the AI. During the prototyping session, participants designed interface elements addressing these requirements and discussed their trade-offs and personal priorities.

The themes and requirements we identified in our workshops showed clear parallels to the psychological work design literature, which links task and work environment characteristics to key psychological processes that shape workers' motivation and job satisfaction. Building on these parallels, we integrated our empirical findings with the SMART model of work design [63], which frames Stimulation, Mastery, Agency, Relatedness, and Tolerable demands as core dimensions of good work. From this synthesis, we developed a design framework comprising twelve considerations for designing oversight interfaces that are not only functional but also support overseers' psychological needs, job satisfaction, and long-term engagement [63], thus supporting "good work".

In summary, the main contribution of this work is a design consideration framework for AI oversight interfaces that demonstrates how interface design can help domain experts fulfill their oversight role, meet their psychological needs,

and derive meaning and satisfaction from their oversight task. This framework is grounded in empirical insights from domain experts overseeing an automated grading system, the user requirements we identified, and concrete examples of interface solutions. By integrating these findings with the SMART model of work design, the framework is broadly applicable to domains beyond grading, such as medical diagnosis and treatment planning, juristical decision-making, asylum evaluations, and many others, where human oversight of increasingly autonomous AI will become essential for responsible AI implementation. Designing effective oversight interfaces is therefore critical not only to prevent errors and support performance but also to ensure that humans remain engaged, motivated, and satisfied in their evolving roles alongside AI.

2 Background & Related Work

2.1 Human Oversight of Al Systems

AI systems are increasingly used to automate or support decision-making in a variety of domains, such as education [66], finance [e.g., 3, 5, 18, 83], healthcare [e.g., 5, 15, 16, 39], or law [e.g., 36, 41, 50, 51, 79] (see Lai et al. [43] for an overview). Human oversight of such AI systems aims to mitigate risks [29, 55]. Broadly speaking, oversight can be considered and implemented at different stages of a systems' lifecycle (e.g. at design time, at run-time, or inspection time [75]) and at different organizational levels (e.g. human oversight or institutional oversight [35, 47]). In this paper, we focus on human oversight of AI at system run-time, as it is required more and more by policies and legislations [35] This includes both, monitoring of the running system to detect failures, unsafe behavior, or biased outcomes [29, 44, 75], as well as intervening in the automated process, such as overwriting decisions or changing inputs, but could also involve delegating the intervention to other stakeholders [75]. Since monitoring and intervention require expert judgment, the role of a human oversight person is typically taken by domain experts, who have the necessary expertise and background to judge the accuracy of AI outputs, or by oversight experts, who are trained to detect AI failures [75].

Sterz et al. [75] argue that an oversight person is effective if and only if four requirements are met: causal power, epistemic access, self-control, and fitting intentions. In other words, an effective oversight person has to have the means to intervene to mitigate risks (causal power), needs enough knowledge to recognize and mitigate them (epistemic access), needs to be in charge of their own doing (e.g., they are not fatigued or drunk) (self-control), and needs to be motivated to do their job properly (fitting intentions). While these requirements for the effectiveness of human oversight need to be supported by its general sociotechnical conditions [45, 47, 75], in this paper, we are particularly interested in how the *user interface* can support effective human oversight of AI systems.

In this regard, research has mostly been done in the broader area of human-AI decision-making, where research has focused on providing the user with a sufficient understanding of the AI system. In particular, the field of XAI (see Ali et al. [2], Langer et al. [46], Longo et al. [52] for reviews of XAI methods for AI-assisted decision-making) has made great efforts to enhance human understanding of the AI [37, 56, 78] to improve the joint human-AI decision-making performance. Furthermore, providing explanations to the user can support their learning process [14, 32]. However, AI-based automation can also lead to unintended negative effects on performance when humans are the final decision makers, such as over-reliance on AI, decreases in situational awareness, or de-skilling [27]. Therefore, recent work has moved beyond automation to explore more active roles for humans in the decision-making process [20, 23, 33, 53, 57, 67], as well as different forms of AI support at various stages of the decision-making process [54, 84].

Instead of decreasing automation, the goal of this paper is to explore how we can utilize the technical advances made in the field of AI decision-making systems while ensuring that humans can perform effective oversight of these

systems to mitigate critical failures or discrimination. Beyond AI, there is large body of work on human oversight of automation, where a main concern is the trade-off between automation bias and doing everything themselves [59]. The higher the level of automation of a system, the more supervisory the humans' role has to be [60]. To support the user in this role, empirical research has focused primarily on cognitive aspects, such as maintaining situational awareness [22, 70], or on technological features that contribute to oversight performance [28].

However, effective oversight requires not only understanding ("epistemic access") and control ("causal power"), but also sustained motivation and engagement: oversight personnel must be willing and able to perform their role diligently and attentively (i.e., maintain "fitting intentions" [75]). This points to an overlooked but critical dimension of AI oversight: its work design.

2.2 Psychological Processes at Work: Individual Needs and Motivation

In terms of a human-centered perspective on work tasks more broadly — beyond human oversight of AI — research has long emphasized that individuals' motivation, performance, and satisfaction are deeply conditioned by the extent to which their psychological needs are fulfilled. Self-Determination Theory (SDT) provides a well-established theory for understanding the sources of human motivation [7, 68, 69]. According to SDT, human motivation is driven by the satisfaction of three fundamental psychological needs: autonomy, competence, and relatedness. When these needs are met, individuals are more likely to be intrinsically motivated to engage in their work tasks [19, 74, 76, 77].

Beyond SDT, the work design literature has emphasized how specific work characteristics can fulfill fundamental psychological needs and thus foster motivation, performance, and satisfaction. Work design refers to "the content and organization of one's work tasks, activities, relationships, and responsibilities" [61, p. 662]. Recently, researchers have integrated decades of research on work design into an overarching model called SMART which describes five work characteristics that influence psychological needs and long-term job satisfaction (Stimulating, Mastery, Autonomous, Relational, and Tolerable) [63]. In contrast to SDT, which outlines *that* people have fundamental needs and what the consequences of these needs are, SMART emphasizes more closely *how* work characteristics can enable need satisfaction.

Although the general work design literature provides extensive insights into how to meet employees' needs and design for good work, including in sociotechnical systems [62], only a few studies have addressed these topics in the context of user interfaces for oversight of AI systems [30]. Yet, work design, psychological need satisfaction, and motivation can be considered highly relevant in the context of oversight of AI [75]. In this paper, we thus draw on the work design literature, specifically the SMART model [63] to formulate specific design considerations for the design of oversight interfaces that mediate the structure, content, and organization of oversight work to support both effective and motivational human oversight.

2.3 Human-centered AI

At a broader level, the core ideas behind the importance of work characteristics and their contribution to motivation and satisfaction also connect to the field of human-centered AI, which emphasizes designing AI systems that respect and support human values, needs, and capacities [71]. Within HCI, psychological processes, particularly human motivation and autonomy, have long been recognized as central to effective interface design [8, 31, 38]. Beyond informing design guidelines, recent research has also examined how specific interface features influence these psychological factors [9, 10, 13]. With the advancements of AI technologies and the growing interactions between humans and AI-based systems, the fields of human-AI interaction [65, 81, 82], human-centered AI [17, 58, 72], and human-centered

XAI [25, 26, 49] emerged. Central to these fields is the recognition that users' needs, values, and psychological well-being must guide the design process, leading to design frameworks and guidelines that foreground these aspects [see 85, for an overview] and to a growing emphasis on participatory design approaches for human-AI interaction [24]. For example, Liao et al. [48] collaborated with UX and design practitioners to create an XAI question bank that captures user needs for explainability and guides designers in addressing them. Similarly, Weitz et al. [80] explored user requirements in social assessment tasks through a co-design workshop with unemployment consultants, while Kim et al. [42] examined end-users' needs and perceptions when using a real-world bird identification app. Together, these studies emphasize that understanding stakeholder needs is fundamental to interface design.

Therefore, in our work we opted to conduct co-design workshops with domain experts to directly elicit their perspectives, needs, and priorities when overseeing AI systems. This participatory approach allowed us to ground our design considerations not only in the broader work design literature but also in the experiences of those who are supposed to perform the oversight work, ensuring that the resulting interface concepts not only enhance effectiveness but also support motivation, autonomy, and psychological well-being, core goals of human-centered AI. In doing so, we extend more general design guidelines for human-AI interaction [4, 6, 34] by addressing the specific psychological and motivational challenges of human oversight roles.

3 Method

The goal of our research is to identify which aspects we should consider when designing oversight interfaces to help domain experts be effective, motivated and engaged in overseeing AI systems. To this end, we adopted a participatory design approach and complemented it with theoretical insights from work design to interpret and structure our empirical findings into a general design consideration framework.

We first conducted four co-design workshops, spanning two sessions each, with domain experts from different backgrounds (i.e., computer science, psychology) which we describe in section 4. With these workshops, we aimed to understand how domain experts approach and experience the oversight of an AI system that automates a task they previously did themselves, what requirements and needs they have for performing this oversight effectively, and how these could be addressed by a user interface. During the workshop, we first tasked participants to oversee an AI-based grading system for student tests and reflect on their own approaches and experiences (subsection 4.3). Together, we then discussed their needs and requirements for an effective and motivational oversight interface. In a second session, participants collaboratively prototyped concrete interface designs, critically engaging with these requirements to explore how they could be addressed in practice. (subsection 4.4). By actively involving domain experts, we ensured that their perspectives and expertise directly shaped the identification of needs and the design of potential interface solutions.

In a second step, we integrated the empirical findings from our workshop with the SMART model of work design [63], which we introduced in subsection 2.2. In section 5, we reflect on how our empirical findings on user requirements and interface prototypes relate to each of the five higher-order work characteristics proposed by SMART (Stimulating, Mastery, Autonomous, Relational, and Tolerable work characteristics) and the psychological processes they condition, and how the theory of SMART can further inform the design of oversight interfaces. By discussing our empirical findings within the scope of each of SMART's work characteristics, we identify a set of specific *design considerations* for developing oversight interfaces.

4 Co-Design Workshop

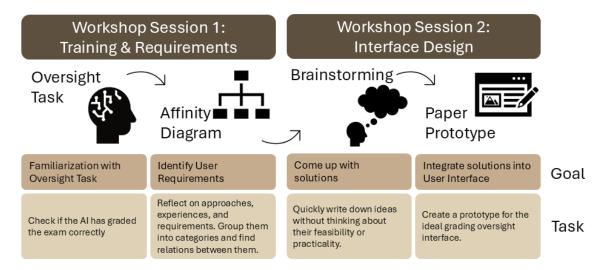


Fig. 1. Overview of the workshop design. We performed two sessions with each group, which both consisted of two tasks with different goals.

We performed four co-design workshops with domain experts from different backgrounds (i.e., computer science, psychology). Each workshop consisted of two two-hour sessions that aimed at different goals (see Figure 1). In the first session, participants conducted the oversight task themselves, and then individually reflected on their approaches and experiences (supported by guiding questions, see section C.1). With the whole group they then created an affinity diagram guided by the workshop moderator (see section C.1) to identify their requirements and needs for support during the oversight process. In the second session, participants brainstormed and selected solutions for the user requirements they identified in the first session, first individually, then together. The groups then created a paper mock-up of their ideal grading oversight interface. Detailed instructions for every part of the workshop are provided in Appendix C.

The oversight task was to oversee an autonomous system that grades student tests. Therefore, participants received a large number of tests (i.e., two exercises with 41 answers each) and had 20 minutes to check if the AI graded them correctly. They received a printed-out sample solution for the exercises, and the students' answers, including the points given by the AI (see Appendix A for an example task). For the task, we used real questions from the 'Programming 1' and 'Work and Organizational Psychology' lectures at the authors' home institutions. In the first case, student answers to exercise questions were generated by the OpenAI o4-mini model. In the second case, responses to the exam questions were drawn as random samples from anonymized student exams. The points were assigned using the OpenAI o4-mini model, which was prompted with the grading scheme. See Appendix B for detailed information about how the prompts were structured. Participants were told that due to the examination regulations of the university, a human oversight person was required to check if the autonomous system performed valid point assignments and to identify possible errors. They were instructed to familiarize themselves with the task and think about different approaches they would use if they had to perform this task in the future. For detailed instructions, see section C.1. The workshop was approved by the first author's institution's ethical review board.

4.1 Participants & Collected Data

The study was performed in-person with four groups of three participants (see Table 1); i.e., N=12 (4 Male, 8 Female, age: 20-32, M=25.5, SD=4.72). The participants were all experts in the task, meaning they had prior grading experience and expertise in the respective domain (i.e., psychology or computer science). We decided to perform the task with computer scientists and psychologists to collect a broad range of ideas, approaches, and user requirements, and to make our findings more generalizable. In a pre-workshop questionnaire, participants indicated that they have not used AI support for grading student exams or assignments before (1-2 on a 5-point Likert scale, M=1, SD=0.39). Participants were recruited by reaching out to researchers in psychology and former tutors of the 'Programming 1' lecture at the authors' institutions.

Group ID	Age	Expertise	Material
CS1	21-23	CS Bachelor and Master Students	Programming 1 Test
CS2	20-22	CS Bachelor Students	Programming 1 Test
Psy1	30-32	Psychology PhD Students and Postdoc	W&O Psychology Exam
Psy2	28-31	Psychology PhD Students	W&O Psychology Exam

Table 1. Demographics of the different workshop groups

The pre-workshop questionnaire included items on participant demographics, their grading experience, and attitude towards AI support tools. Participants created an affinity diagram of their approaches, experiences, and user requirements when performing the oversight task, collected ideas about how a user interface could support them in this task, and created a paper mock-up of a user interface for the oversight task. The participants' answers to the pre-workshop questionnaire, the affinity diagrams, the ideas collected for a user interface, and pictures of the final paper mock-ups are provided in the supplementary materials. All workshops were video- and audio-recorded. They were performed in English (CS2) or German (CS1, Psy1, Psy2). For this paper, participants' quotes were translated and shortened for clarity.

4.2 Data Analysis Approach

To analyze the data collected during the workshops, we employed reflexive thematic analysis [11, 12]. It supports inductive, flexible theme development while allowing the researchers' perspectives and expertise to play a meaningful role in the analysis. We did not transcribe the video and audio recordings. Instead, coding was performed directly on the video materials to preserve contextual and nonverbal cues (e.g., the spatial layout of affinity diagrams, gestures, and emphasis in group discussions). Three researchers (all co-authors of this paper) conducted the analysis. Coders 1 and 2 each analyzed the workshop(s) they had facilitated. A third coder, who had not facilitated any session, coded all workshops independently to enhance interpretive diversity and support analytical rigor. All authors then participated in identifying and refining themes from the codes. The qualitative analysis proceeded through the following steps:

- 1. Familiarization. Each coder reviewed their assigned workshop materials fully to immerse themselves in the data.
- 2. *Initial Coding*. The coders developed descriptive and interpretive codes that closely reflected the participants' own language and captured their relevant thoughts.
- 3. *Theme Development.* The codes were grouped into broader conceptual categories, i.e., themes, to identify recurring patterns and points of convergence related to the research questions.
- 4. *Theme Refinement.* Themes were iteratively reviewed to ensure internal coherence and clear distinction. During this process, some themes were merged, split, or redefined to more accurately represent the underlying data.

5. *Theme Naming.* Each theme was given a descriptive name that reflected its central concept and function within the broader thematic structure and the researchers' higher-level interpretation of recurring ideas in the data.

6. *Relationship Construction*. Coders examined the relations among themes within and across research questions. This included identifying conceptual overlaps, as well as inferred cause-effect relationships.

These stages were informed by both the researchers' close engagement with the data and their interpretive synthesis, meaning the process of drawing meaning from patterns across the data based on analytical judgment, contextual understanding, and iterative reflection. Importantly, this process was also guided by the researchers' prior empirical and theoretical knowledge, particularly from the previously discussed literature. Reflexive thematic analysis explicitly allows for such theoretically-informed interpretation and positions researcher subjectivity as a resource rather than a limitation. Through our analysis approach with three independent researchers from diverse backgrounds (Computer Science & Philosophy (Coder 1) and Work & Organizational Psychology (Coder 2,3)), we ensured interpretive diversity and analytical rigor. Finally, the coding team synthesized their results collaboratively. Through discussion, the coders compared and reconciled interpretations, refined the thematic structure, and jointly confirmed the relationships between themes. The final thematic framework is presented in Figure 2. In the following, we first present results on the approaches and experiences of the domain experts (subsection 4.3) and then on the identified user requirements and their implementation in UI prototypes (subsection 4.4).

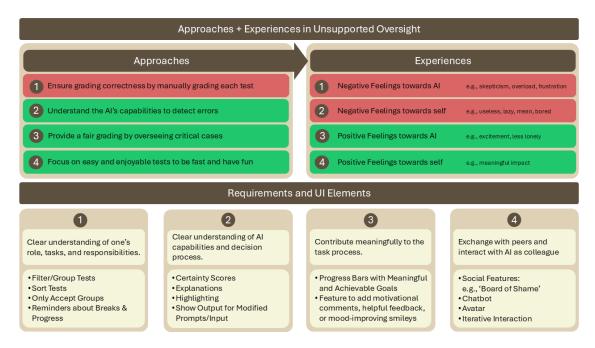


Fig. 2. Overview of the main themes identified in our thematic analysis.

4.3 Results: Initial Approaches and Experiences of Domain experts

The most prevalent finding across all groups was that participants were drawn to *perform the AI's grading task* instead of fulfilling their role as an oversight person. This approach undermined efficiency gains and led to mainly *negative*

feelings when performing the task. Approaches that participants perceived as more positive included detecting unjustly awarded points by understanding the AI's strengths and weaknesses, ensuring fair grading of borderline cases, and focusing on tests that were easy and enjoyable to grade. For each of the approaches, we discuss the underlying goal and the strategy the participants adopted to achieve it below, as well as the positive and negative experiences participants reported with these approaches.

4.3.1 Approaches.

Approach 1. Ensure grading correctness by manually grading each test.

Most participants initially did not approach the task as an oversight task, but as a grading task, which was characterized by the participants' goal to check all of the students' tests (CS1, CS2, Psy1, Psy2) by grading them manually and comparing their grade to the AI's grade (CS1, CS2, Psy1, Psy2). Reflecting on this initial approach, participants pointed out that their goal was not achievable and the strategy, therefore, was not efficient (CS1, CS2, Psy1, Psy2; Psy 1, P1: "I didn't really find it efficient. You still had to read everything, you still had to assess everything and then compare that with your own grading").

Approach 2. Detect unjustly awarded points by understanding the AI's strengths and weaknesses.

After reflecting on their initial approach, participants described that the goal to check only a subset of the tests was more achievable than aiming for all tests (CS1, CS2, Psy1, Psy2) and that strategies that included skimming the answers much more superficially, as compared to when doing the grading manually, are more efficient (Psy1). Therefore, it was proposed to aim to find unjustly awarded points (CS1, Psy1) by understanding the strengths and weaknesses of the AI system and focus only on a subset of tests, which they judged especially important for their task (CS1, CS2, Psy1, Psy2). Participants deemed extreme cases important, such as tests with zero or full points or unique answers, since they expected the AI to perform worse in these cases (CS1, CS2, Psy1, Psy2).

Approach 3. Provide fair grading by overseeing critical cases.

Another goal of the participants was to provide fair grades (CS1, CS2, Psy1) by focusing on subsets of tests, which they judged especially important (CS1, CS2, Psy1, Psy2). This resulted in the approach to identify critical cases in which the AI assigned too few points, causing the student to fail the test (CS1).

Approach 4. Perform the task fast and have fun by overseeing tests that are easy and enjoyable to grade.

Lastly, participants discussed the goal of performing the task fast and having fun (CS1, CS2, Psy1, Psy2). This goal was pursued by focusing on the tasks that were easiest (CS1, CS2) or most fun to check (CS2).

4.3.2 Experiences.

Experience 1. Negative feelings towards AI.

Most negative experiences related to their feeling that human-AI collaboration was not beneficial for this task. Participants felt skeptical of AI (Psy1), unsupported (CS1, CS2, Psy1, Psy2), or like doing extra work (CS1, Psy1, Psy2). They felt negative overall (CS1, Psy1), frustrated (CS1, Psy1), stressed (Psy2), confused (Psy1), uncertain (Psy1), and biased (CS1, CS2, Psy1). One group (CS1) linked their negative experience to their perceived responsibility for possible errors, since they felt obligated towards the students that humans should assess their tests: "Somebody should be responsible for the grade, and we can't say it is the AI" (CS1, Participant 2).

Experience 2. Negative feelings towards self.

Performing the oversight task also caused negative feelings about themselves. One participant (CS1, P1) described that they focused on tests, where the AI gave full points, which seemed most efficient, but made them feel bad about themselves: "In the end, if you decrease points from people, you think, if I did not change it, they would have more points and I screwed them up. That's the feeling that resonates."(CS1, P1). Other participants felt useless performing the task (CS2), which made them feel lazy (CS1), bored (CS2, Psy1), or tired (CS2, Psy1).

Experience 3. Positive feelings towards AI support.

Despite the rather negative experiences, some participants felt hopeful or excited based on the prospect of saving time and energy (Psy1). One group (Psy1) reflected that the valence of their experiences depended on the specific cases they were overseeing. For example, some participants judged the AI's grading as very good for one of the two exercises they received, but did not understand its recommendation for the other (Psy1, P2: "There is definitely satisfaction [with the system] but kind of limited to specific questions").

Experience 4. Positive feelings towards self.

Furthermore, one participant felt less lonely when performing the task and safer in their decision-making, because they could share the responsibility for the outcome (Psy2), and another approached the task intending to have a meaningful impact (CS1, P2). Aiming to prevent students from failing the test due to an AI error, they had a more positive experience than the other participants in their group: "Too many points do not hurt the student, but rather if the grading is done too strictly; I think it is somehow wrong [...] to always decrease their gradings" (CS1, P2).

4.4 Results: User Requirements and Supporting UI Elements

Reflecting on the oversight task, the participants described different user requirements for a good oversight interface. We found four main themes when discussing their user requirements and needs: role understanding, AI understanding, meaningfulness, and relational aspects. Participants also came up with UI elements, that support the interface to meet their requirements.

4.4.1 Role Understanding: Clear understanding of one's role, the associated responsibilities, and tasks. Participants desired a clear understanding of their role, which should be reflected in the user interface's affordances. When performing the oversight task, they wished for a clear understanding of their responsibilities (CS1, Psy1, Psy2), since the task itself did not clarify the responsibility distribution, resulting in varying perceptions. Furthermore, they wanted the interface to clearly represent what their task is to avoid misconceptions based on prior experiences: (Psy1, Psy2; Psy2, P2: "I was aware that I should oversee the AI, but because of the formatting of the answers, I reverted to old habits of grading exams as I did it in the past").

Therefore, two groups actively decided not to include an accept/reject decision for every single test, but only a button to finish the task (CS2, Psy2), and one group separated the task of detecting errors from fixing them (Psy2). All workshop groups tried to focus on subsets of tests that are important to oversee by grouping or sorting the tests according to different criteria (CS1, CS2, Psy1, Psy2) (see Figure 3). Additionally, participants aimed to remind users about their oversight role and support them in not grading too many tests themselves by including reminders about taking breaks (Psy2) or finishing the task after a certain time (CS1, CS2).

4.4.2 Al Understanding: Clear understanding of Al capabilities and decision process. Participants expressed their desire to understand the Al's capabilities and its decision process and requested system features to support them (see Figure 4)



(a) Grouping of special cases with (b) Function to filter tests accord-(c) Selection for different kinds of progress bar - CS1 ing to different criteria - CS2 tasks or answers - Psy2

Fig. 3. UI Elements that participants included to filter and group the student answers.

(CS1, CS2, Psy1, Psy2). Participants wished for information about the AI's strengths and weaknesses, such as certainty/confidence scores (CS1, CS2, Psy2), explanations (CS2), or accuracies (Psy2). Further, they desired to understand the AI's reasoning and decision-making process (Psy2, P2: "It should be clear — the steps behind — how the AI came to the decision"; P1: "We mentioned that a lot: for us transparency is very important"). The UI should show (parts of) the process that led to the decision by highlighting decisive parts of the answer (CS1, Psy1, Psy2), providing a general grading policy (CS1), assessing the answer based on various prompts (CS1), or providing an improved student answer (CS1). Interestingly, they hypothesized that their little understanding of the AI also decreased their role understanding: "I think the difficulty of focusing on overseeing the system is that you don't have much information about the system. Otherwise, this shift would have been easier." (Psy2, P33).



Fig. 4. UI Elements that participants included to better understand the AI capabilities and its decision process.

4.4.3 Meaningfulness: Contribute meaningfully to the task process. Participants wanted to perform a meaningful task. In group CS1, participants found that pursuing a meaningful goal (i.e., detecting correct features of answers that the AI missed) made them feel satisfied. They saw opportunities for meaningful additions to AI grading that would support students, such as motivational comments, helpful feedback, or mood-enhancing smileys (CS1, CS2, Psy2). However, meaningfulness for them was (at least partially) constituted by their role understanding, performance, and control over the AI and outcomes. In their interface prototypes, meaningful goals were mainly apparent through progress bars, which would only benefit them if they aimed for meaningful goals (CS1, CS2, Psy1, Psy2). Therefore, two groups included progress bars towards one or multiple small sub-goals (CS1, CS2), and one group (Psy2) removed the progress bar again. But also providing confidence levels for the grading, they would "start with the things that AI has the lowest confidence in—where they need me most" (Psy1, P2).

4.4.4 Relational Aspects: Exchange with peers and interact with AI as colleague. A common theme that occurred in all workshop groups was the desire for social interaction when performing the oversight task. This social interaction could take different forms: with other oversight people, the students, or the AI. In their prototypes, participants included features that supported them to meet their relational needs (see Figure 5).

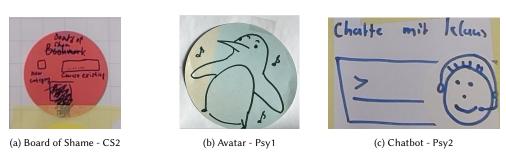


Fig. 5. UI Elements that participants included in their prototypes to meet their relational requirements.

Participants described the desire to connect with other oversight people by exchanging about different student answers they encountered (CS1, CS2). One group (CS2) added a 'Board of Shame' (Figure 5a) to the interface where they could share funny or creative student answers with other tutors: "If the button was not called 'Bookmark' but 'Board of Shame', [...] I would for sure bookmark every weird edge-case I encountered. That would be more fun." (CS2, P3) [...] "that would be extremely motivating." (CS2, P1). Furthermore, social interactions were important for deciding which information to display in the interface (CS1).

Participants wanted to have human-like interactions with AI to feel connected (Psy2) or experience the task as more enjoyable and fun (Psy1, Psy2). Group Psy2 added a chatbot with a human-like name and appearance, while Group Psy1 used the likeness of a penguin, which dances or smiles when good student answers were displayed on the screen. (Figure 5c). They wanted the AI to appear as a colleague with which they can communicate and discuss to understand its recommendations better, improve the grading, and share responsibility (CS1, Psy1, Psy2). Notably, group Psy1 stressed how this relational need and its fulfillment were prone to inter-individual differences.

5 Theoretical Integration and Design Considerations

In the following, we reflect on how our empirical findings on user requirements and interface prototypes relate to each of the five higher-order work characteristics proposed by SMART (Stimulating, Mastery, Autonomous, Relational, and Tolerable work characteristics) and the mediating psychological states (e.g., work meaningfulness, fulfillment of relational needs) that ultimately shape job satisfaction. Although SMART primarily addresses work design at the job level, these characteristics also apply to user interfaces and technologies through which work is performed, influencing the same psychological processes. Considering these characteristics at the interface design level could thus help to design user interfaces that are not only functional but contribute to job satisfaction and long-term performance [63]. By discussing our empirical findings within the scope of each of SMART's work characteristics, we identify a set of design considerations for developing oversight interfaces, summarized in Table 2. These design considerations are not prescriptive guidelines but rather points for reflection, acknowledging that enhancing one characteristic may trade off with another or alter the oversight task. Table 2 further illustrates how participants expressed each consideration in their interface prototypes.

Design Consideration for Oversight Interfaces	Example from Co-Design Workshop and Grading Task	Psychological Processes
Stimulating Interface Characteristics Skill Variety. Consider how the interface can encourage variety in the skills applied during oversight.	The interface offers the user a varied set of tests to check, including different exercise types, answer lengths, or different types of student mistakes, to make the oversight task less monotonous and more stimulating.	Challenge Appraisals and Work Meaningfulness
Meaningfulness of oversight. Consider how the user can contribute to the automated task in a way that is meaningful to them and reflect that in the interface.	The interface displays borderline cases in a dedicated area to help oversight personnel ensure that no student unjustly fails the test.	
Mastery Interface Characteristics Understanding of Al. Consider which information about the Al's capabilities and decision process supports the user in their oversight task. Understanding of Oversight Role. Consider how UI elements influence the user's role perception. Feedback. Consider how the user interface can indicate task progress and the impact of interventions on the final	For each type of exercise, the interface displays an average certainty value of the Al grading. This could help the user understand the Al's overall strengths and weaknesses better than showing certainties for each individual test. The interface has a "change grading" button but no "accept" button to prevent the user from re-grading each individual test. The interface displays statistics and metrics about the grading, such as point distributions. If the oversight person adjusts the grading scheme or lets the Al reevaluate the tests, they	Challenge Appraisals and Activated Negative Affect
Autonomous Interface Characteristics Timing Autonomy. Depending on the timing constraints, consider how the interface can give the user autonomy to organize their responsibilities. Method Autonomy. Depending on the oversight constraints, consider how the interface can give the user autonomy about how they monitor the Al and how they intervene.	immediately see the overall impact of their intervention. The interface offers the user the possibility to "save" individual tests to a dedicated area from which they can be re-graded at any time. This is possible since grading is an oversight task without time-critical interventions. The interface offers sorting and grouping options to allow the user to choose and prioritize which tests to review.	Work Meaningfulness
Relational Interface Characteristics Relationship to peers. Consider how the interface can support the user in relating to other oversight persons. Relationship to affected individuals. Consider mechanisms that help users to recognize and positively shape the impact of their oversight on others. Relationship to AI. Consider how the appearance and interaction with the overseen AI contribute to the user's relational needs but might affect oversight performance.	The interface provides the option to share interesting answers with other people involved in the oversight, which they might discuss outside the oversight process. The interface supports allows the user to add motivational comments or constructive feedback to the automated grading. The interface lets the user exchange with a human-like Al about the decisions, which lets the oversight person connect more easily with them, but may increase bias.	Work Meaningfulness and Fulfillment of Relational Needs
Tolerable Interface Characteristics Role Overload. Consider how the interface can present the requirements of the oversight task in a manageable way. Role Conflicts. Consider how the interface can clearly separate the tasks and responsibilities of the user from those of the Al.	The interface provides metrics on different exercises and overall grading, which help the user manage their resources efficiently. The interface allows users to improve the Al's grading solely by requesting a reevaluation, which reduces role conflict, as users do not grade tests themselves.	Activated Negative Affect

Table 2. Considerations for designing oversight interfaces, grouped by their mapping to the Work Characteristics proposed by the SMART model of work design [63] and links to psychological processes as proposed by SMART. Each consideration is illustrated by an example from our co-design workshop that demonstrates how participants considered each point in their own prototypes.

5.1 Stimulating work characteristics

Stimulating work is characterized by task variety and the opportunity to draw on different skills and requirements for problem-solving and information processing. When these features are absent, work tends to be experienced as

monotonous and demotivating. In our study, oversight resembled the repetitive, low-level tasks that SMART links to reduced stimulation, with participants describing re-checking straightforward AI outputs as boring and tiring. Oversight appeared more engaging when participants focused on ambiguous, extreme, or idiosyncratic cases, which SMART suggests may enhance cognitive challenge and stimulation.

Oversight interfaces should thus contribute to oversight persons being able to use different skills for problem-solving and information processing. Participants in our workshop recognized that some tests were more stimulating to check than others and added interface elements to highlight them, for example, by displaying outliers or ambiguous answers in a dedicated space, or by enabling sorting based on AI certainty or received points (see Figure 3).

C1. Skill Variety. Consider how the interface can encourage variety in the skills required for oversight.

The pronounced concern among participants to contribute meaningfully to the process (see subsubsection 4.4.3) and to perform the task well, suggests that the task was not experienced as sufficiently stimulating on its own. According to SMART, insufficient stimulation diminishes perceived meaningfulness, which in turn lowers motivation and satisfaction, as evidenced by the participants' negative experiences (subsubsection 4.3.2).

Thus, designers should consider how the user interface can support domain experts in contributing meaningfully by overseeing the automated task. Participants saw meaning in preventing students from unjustly failing the exam. Thus, they integrated filtering and sorting features to focus on students who were just below the passing threshold.

C2. Meaningfulness of oversight. Consider how the user can contribute to the automated task in a way that is meaningful to them and reflect that in the interface.

5.2 Mastery work characteristics

According to SMART, workers require an understanding of what their tasks are, how these tasks fit into the broader process, and how well they are performing. The salience of this issue was reflected in the themes of role understanding (subsubsection 4.4.1) and AI understanding (subsubsection 4.4.2). Participants were often uncertain whether they should grade or oversee, unclear about their responsibilities, and lacked insight into the AI's decision-making, which might have helped them better understand their role and contribution. Uncertainty about their role undermines mastery and fosters negative affect, which aligned with participants' negative experiences of uselessness and overload (subsubsection 4.3.2).

Designers should consider UI elements that support understanding and monitoring of AI decisions, clarify the oversight task through intervention mechanisms, and provide feedback on task progress and the impact of human oversight.

In their interface prototypes, participants included features like confidence scores and AI output explanations. One participant suggested displaying the AI's grading policy rather than individual test results to promote higher-level oversight, expecting this would enhance understanding of both the AI's capabilities and their own oversight responsibilities. Some observed that the option to directly modify grades encouraged them to take over the grading task, leading to its removal from the interface. Instead, one group proposed allowing the AI to re-evaluate tests based on additional input or instructions. Most participants included progress bars, indicating a need for UI support in tracking performance. Additional feedback could highlight the impact of their interventions, for example, metrics showing how many students passed due to their oversight, how many AI errors they identified, or how re-evaluations they triggered changed the grading.

C3. Understanding of AI. Consider which information about the AI's capabilities and decision process supports the user in their oversight task.

- C4. Understanding of Oversight Role. Consider how UI elements influence the user's role perception.
- **C5. Feedback.** Consider how the user interface can indicate task progress and the impact of interventions on the final output or the AI system.

5.3 Autonomous work characteristics

According to SMART, autonomy refers to perceived control over timing, methods, and decisions, enabling ownership, which in turn promotes a sense of meaningful work. Autonomy emerged as a recurring tension in the oversight task. Because automated grading is not time-sensitive, participants generally had high timing autonomy and rarely operationalized it. While some expressed a desire to time their actions freely, they chose to limit this autonomy to maintain clearer role boundaries. This raises the question of whether UI support for timing autonomy may be more critical in time-sensitive tasks. While participants had method autonomy in how to perform oversight, the accompanying ambiguity about decision rights and accountability weakened their sense of control and ownership (see subsection 5.2). This helps explain participants' frustration and doubts about their role's purpose, suggesting that method autonomy alone is insufficient for meaningful work without clearly defined scope and limits.

These tensions highlight the need for designers to provide users with an appropriate degree of autonomy, clearly communicated through the interface. For work to feel meaningful, users require both autonomy and a clear understanding of their role. Designers should define which freedoms align with the user's role and task, make role boundaries visible, and offer meaningful choices within them. While too many options can overwhelm users, too little autonomy can leave them feeling restricted and ineffective.

One group discussed that oversight involved both monitoring the AI's grading and correcting errors. To avoid reverting to manual grading, they preferred postponing corrections and felt that limiting their ability to intervene directly would enhance role clarity by reducing reminders of their previous grading role. Importantly, they valued timing autonomy, but omitted it in favor of preserving role clarity. Participants exercised method autonomy primarily by choosing which subset of student answers to oversee, but also discussed allowing users to choose their level of oversight, such as reviewing individual tests, the AI's grading policy, or grading metrics. Designers should also consider what types of interventions to support, such as correcting the error themselves, delegating it, or prompting the AI to re-evaluate specific tests.

- **C6. Timing Autonomy.** Depending on the timing constraints, consider how the interface can give the user autonomy to organize their responsibilities..
- **C7. Method Autonomy.** Depending on the oversight constraints, consider how the interface can give the user autonomy about how they monitor the AI and how they intervene.

5.4 Relational work characteristics

Relational work characteristics support the need for relatedness by enabling social interaction and a sense of contributing to others within a broader context. Since the oversight task offers limited opportunities for relational fulfillment, participants expressed their desire for exchanges with peers about unusual or noteworthy student answers and connecting with students through meaningful feedback, motivational comments, or small gestures of recognition. From a SMART perspective, these social and beneficiary-oriented contributions support the psychological need for relatedness.

Interface designers should consider how the UI can support users' relational needs, not only through social features, but also by displaying information that facilitates later peer exchange. Additionally, when the automated task impacts others, the interface should highlight the user's impact on those individuals.

In the workshop, participants discussed ways to exchange with peers about the oversight. One group included a 'Board of Shame' (Figure 5a) in their prototype to share unique or humorous student answers. Others added student identifiers to facilitate discussions about specific solutions, even outside the interface. Participants incorporated commenting features into automated grading to provide meaningful feedback and recognition. Further, they integrated sorting mechanisms and dedicated displays for critical cases to mitigate the risk of students failing due to AI errors.

- **C8. Relationship to peers.** Consider how the interface can support the user in relating to other oversight persons.
- **C9. Relationship to affected individuals.** Consider mechanisms that help users to recognize and positively shape the impact of their oversight on others.

Interestingly, some participants expressed a desire to engage with the AI in relational terms, envisioning it as a collaborator with whom they could "discuss" recommendations, question decisions, and share responsibility, and gave it a human or animal-like appearance (Figure 5). While SMART primarily conceptualizes relational work in human-human terms, these accounts suggest that similar psychological needs may extend to interactions with AI. Therefore, it is essential to consider how users interact with the AI they oversee. Users may satisfy their need for relatedness by engaging with AI in human-like ways, yet this risks bias, over-reliance, or emotional attachment [1, 64, 73]. Designers must therefore consider how to shape AI appearance and interaction so users can meet relational needs without compromising oversight.

C10. Relationship to AI. Consider how the appearance and interaction with the overseen AI contributes to the user's relational needs but might affect oversight performance.

5.5 Tolerable work characteristics

Tolerable work assesses whether work requirements avoid overwhelming workload (low role overload), conflicting task requirements (low role conflict), or clashes with other responsibilities (low work-home conflict). Our participants' experiences indicated role overload and, possibly, role and work-home conflict. Some participants experienced overload when manually grading all student submissions rather than overseeing the AI, a strategy recognized as inefficient but adopted, for example, from a sense of responsibility (subsection 4.3). This approach may have caused role conflict, as participants' self-imposed expectations (ensuring all grading was correct) clashed with their oversight role. The tension was intensified by the perception that responsibility for errors ultimately rested with them, to the extent that they were willing to complete the task in their free time or on weekends, thereby creating tensions between work responsibilities and private life. From a SMART perspective, these accounts illustrate how tolerable demands were not always achieved in the oversight task. Although the task was narrow in scope, attempts to balance efficiency, fairness, and responsibility sometimes made the demands feel difficult to manage.

This balance underscores the need for designers to present user demands in a manageable way through the interface. Rather than overloading users, the interface should help them manage tasks efficiently and clearly distinguish between user and AI responsibilities.

In the workshop, participants added filtering and sorting functions and progress bars to help focus on achievable goals without becoming overwhelmed. Some groups also included reminders showing time spent and progress made

to manage task load. To reduce role conflict and distinguish their oversight role from the Al's grading, participants removed the 'accept' button for individual tests and, in some cases, refrained from making direct corrections.

- **C11. Role Overload.** Consider how the interface can present the requirements of the oversight task in a manageable way.
- **C12. Role Conflicts.** Consider how the interface can clearly separate the tasks and responsibilities of the user from those of the AI.

6 Discussion and Conclusion

This paper investigated effective and motivational human oversight of AI from a user-centered perspective. We analyzed experiences, discussions, and interface prototypes from four co-design workshops where domain experts were tasked to first oversee an automated AI grading system, then discussed their experiences and personal needs for oversight support, and finally reflected on these user requirements by creating concrete interface prototypes that could support them during AI oversight. Our thematic analysis revealed that domain experts often relied on approaches they later judged ineffective, leading to negative feelings towards both the AI and themselves. From participants' discussions, we inferred four user requirements for oversight support. Participants desired a clear understanding of their role, tasks, and responsibilities, as well as of the AI's capabilities and its decision process. They further wanted to contribute meaningfully to the task process, exchange with peers, and interact with the AI as a colleague. Through prototyping, participants critically reflected on these requirements and discussed how to implement them in practice.

Importantly, we found that the identified user requirements aligned closely with the work design literature, which emphasizes the support of users' psychological needs in interface design. Specifically, we integrated our findings with the SMART model of work design [63] and proposed a comprehensive consideration framework for the design of user interfaces that support human oversight of AI. It centers around five general work characteristics, which we apply as characteristics of the interface: Stimulating, Mastery, Autonomous, Relational, and Tolerable interface characteristics. Each characteristic links the user requirements from our workshops with psychological processes found to impact long-term job satisfaction. Being rooted in a general theory of work design and informed by the task-specific outcome of our co-design workshops, we believe that this consideration framework will be applicable to a variety of domains where AI will increasingly be used to automate decision-making processes, such as in legal, medical, or educational domains, and many other.

In the following, we discuss the trade-offs that should be considered when designing for different interface characteristics and how our proposed consideration framework relates to existing requirements for effective human oversight and to broader human-AI interaction guidelines.

6.1 Trade-offs and Task Constraints

Our consideration framework deliberatively avoids prescriptive guidelines, since designing for specific interface characteristics might entails trade-offs that may unintentionally affect other aspects of oversight or depend on the specific task and organizational context. When aiming to enhance particular work characteristics, designers should therefore remain mindful of potential tensions between considerations and carefully balance competing design goals.

For example, designing for a strong *relationship with affected individuals* could increase an oversight person's *role overload* or *role conflict* if they find their responsibilities difficult to fulfill. Also, the quantity and nature of *feedback* should be carefully considered to prevent *role overload* or altering their *role understanding*: if the overseer does not aim

to re-evaluate all decisions, showing information about the number of tests they did not re-evaluate may overwhelm them or even alter their role understanding. Excessive *autonomy* in task execution may similarly induce *role overload* or hinder *role understanding*. Instead, designers may choose to restrict overseers in how they perform the oversight task, which they should, in turn, balance against users' *autonomy* and perception of the *meaningfulness of oversight*.

Interface designers should also consider that some characteristics of SMART were more relevant and observable through our workshops and can be considered for interface design more easily than others. Therefore, it is important to remember that all of these characteristics can be considered at multiple levels — some are easier met at the interface level, while others need to be considered from the task or job level. Sometimes the task itself may negatively affect a specific work characteristic, which can then be accounted for at the interface level, or the other way around. For example, for a mundane task where only very little information needs to be processed and interventions do not require critical thinking, challenge appraisal or work meaningfulness could be addressed through other stimulating interface characteristics or at the job level (e.g., combining oversight with other tasks). Interface designers should also consider that an oversight person's relational needs may be met outside their oversight task. When other work activities strengthen the relationship to peers, relational features in the interface are less critical. However, when such opportunities are scarce, the interface should actively foster relatedness. Different oversight tasks offer, for example, varying levels of feedback. Some provide immediate responses to interventions, while others reveal outcomes only over time. In this case, it is important to consider that feedback can also come from organizational sources, such as supervisors, or through training that enhances understanding of the oversight role and the AI's capabilities. Lastly, work must be tolerable. While interfaces can present workload in manageable ways, supervisors should avoid assigning unmanageable responsibilities. In high-load tasks, preventing role overload by presenting information in a way that supports manageability becomes especially critical.

6.2 Design Considerations for Effective Human Oversight

While our consideration framework is mainly informed by the work design literature and findings from our co-design workshops, it also aligns with theoretical frameworks on effective human oversight. Specifically, Sterz et al. argue that "an oversight person is effective in their human oversight if and only if they have causal power, epistemic access, self-control, and fitting intentions" [75, p.2499]. According to this reasoning, effective oversight requires the ability to influence the system in ways relevant to one's goals (causal power) [75], a requirement reflected in our considerations: an oversight person who identifies with their role perceives their contribution as meaningful. Furthermore, oversight personnel need sufficient knowledge of their decision situation, including the system, its state, potential interventions, and their consequences (epistemic access) [75], which aligns with our considerations for the mastery work characteristics. Sterz et al. further emphasize that effective oversight requires the oversight person to act autonomously and retain their attention (self-control) [75]. Similarly, our considerations highlight the importance of a stimulating task and ensuring the users' autonomy. Finally, the oversight person needs the intention to be effective (fitting intentions) [75]. While designers may not prevent people with conflicting interests or ill intentions from performing ineffective oversight, designing motivating and satisfying experiences could increase the likelihood of appropriate intentions.

Similar to our prior discussion, Sterz et al. argue that technical design (e.g., interface design), individual factors (e.g., training, domain expertise, motivation), and environmental factors (e.g., job design, accountability, time pressure) influence the extent to which these requirements are met. Overall, our considerations align with their requirements for effective human oversight but extend them by integrating work design principles and providing concrete guidance for interface designers.

6.3 Extending Human-Al Interaction Guidelines for Human Oversight

We also compare our design consideration framework with existing guidelines for human-AI interaction design (see [85] for an overview), since human oversight of AI could be considered a specialized form of human-AI interaction. Specifically, we reviewed three popular guidelines published by Google [34], Apple [6], and Microsoft Research [4]. We see an overlap in key aspects of our considerations, including promoting understanding of AI, granting autonomy, providing feedback, and preventing role overload and conflict. However, we also see that our considerations extend on these guidelines in three important ways. First, we focused on the interaction between an AI that automates a task and a human overseeing its output. While participants encountered challenges, they were also optimistic and excited that the AI automated tasks they previously found meaningful and stimulating (Participant 3, Psy1: "In the end, I feel like it is still kind of exciting because maybe it does already kind of set a baseline"). Current guidelines recommend augmenting rather than automating stimulating tasks [34], yet in high-risk or critical work contexts, automation may be justified, for example, to enhance safety. Our considerations emphasize the importance of UI design to support users in understanding their role as an overseer and highlight other opportunities for experiencing work meaningfulness. Second, existing guidelines address role conflict by recommending to adapt AI behavior to align with users' initial behavior [4, 6, 34]. Our workshop revealed that participants initially experienced role conflict but resolved it by including specific UI elements and controls that would help them to adjust their own approach to the task (e.g., removing an Accept button). This underscores that role conflicts can be mitigated through changes in both AI and user behavior. Lastly, current guidelines largely neglect users' relational needs. Our findings highlight the value participants placed on relating to peers and those affected by their actions. These relational needs may be amplified by the specific task setting, suggesting that our considerations offer unique insights for designing oversight interfaces and that further research is needed to explore relational needs in other oversight contexts and human-AI interactions at the workplace.

6.4 Limitations and Future Work

A key limitation of our workshop concerns the characteristics of the oversight task. Domain experts graded a set of tests to evaluate the AI's performance and detect potential errors, with no direct time pressure. While non-time-critical tasks are common in education, law, or healthcare, other domains, such as aviation, involve trained oversight experts or real-time interventions, which introduce different work characteristics. Consequently, the relevance of each consideration depends on the broader work context. For instance, in our grading task, participants had substantial autonomy in timing and method, making interface support for autonomy potentially less critical than in tasks with lower inherent autonomy. Future work should validate and potentially extend our considerations with requirements from other types of tasks.

Our participants had limited experience with automated grading tools and performed the task only briefly at the start of the workshop. In real work settings, oversight occurs over longer periods and repeated sessions, which may reveal additional problems or needs. Nevertheless, we expect our findings to generalize to longer interactions with AI-based decision-support systems, since prior research suggests that negative experiences and unmet psychological needs may intensify over time [30].

Still, future work should investigate the user requirements of oversight personnel through longitudinal studies, across different domains, and also study individuals who regularly perform oversight. Another promising direction is to further examine the trade-offs between our considerations and provide more guidance on how to design good work for

oversight personnel. Most importantly, though, more empirical work will be needed to evaluate the impact of specific considerations on users' psychological processes and their satisfaction with the oversight task.

Acknowledgments

This work was partially funded by the DFG grant 389792660 as part of TRR 248 CPEC – Center for Perspicuous Computing.

7 GenAl Usage Disclosure

The authors of this paper used generative AI for editing the manuscript, improving the quality of existing text, and creating material used in the co-design workshop. In section 4, we describe how the materials were generated and provide information about the prompts we used in Appendix B.

References

- Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2025. All Too Human? Mapping and Mitigating the Risks from Anthropomorphic AI. In Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24). AAAI Press, San Jose, California, USA 13-26
- [2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion 99 (Nov. 2023), 101805. doi:10.1016/j.inffus.2023.101805
- [3] Yasmeen Alufaisan, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. Does Explainable Artificial Intelligence Improve Human Decision-Making? Proceedings of the AAAI Conference on Artificial Intelligence 35, 8 (May 2021), 6618–6626. doi:10.1609/aaai.v35i8. 16819 Number: 8.
- [4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk, 1–13. doi:10.1145/3290605.3300233
- [5] Ksenia Appelganc, Tobias Rieger, Eileen Roesler, and Dietrich Manzey. 2022. How Much Reliability Is Enough? A Context-Specific View on Human Interaction With (Artificial) Agents From Different Perspectives. Journal of Cognitive Engineering and Decision Making 16, 4 (Dec. 2022), 207–221. doi:10.1177/15553434221104615 Publisher: SAGE Publications.
- $[6] \ \ Apple.\ 2023.\ Human\ Interface\ Guidelines\ for\ Machine\ Learning.\ https://developer.apple.com/design/human-interface-guidelines/machine-learning$
- [7] Nick Ballou, Sebastian Deterding, April Tyack, Elisa D Mekler, Rafael A Calvo, Dorian Peters, Gabriela Villalobos-Zúñiga, and Selen Turkay.
 2022. Self-Determination Theory in HCI: Shaping a Research Agenda. In Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 113, 6 pages. doi:10.1145/3491101.3503702
- [8] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and Accountability: Human Considerations in Context-Aware Systems. Human-Computer Interaction 16, 2-4 (Dec. 2001), 193–212. doi:10.1207/S15327051HCI16234_05 Publisher: Taylor & Francis _eprint: https://doi.org/10.1207/S15327051HCI16234_05.
- [9] Daniel Bennett and Elisa D. Mekler. 2024. Beyond Intrinsic Motivation: The Role of Autonomous Motivation in User Experience. ACM Transactions on Computer-Human Interaction 31, 5 (Oct. 2024), 1–41. doi:10.1145/3689044
- [10] Dan Bennett, Oussama Metatla, Anne Roudaut, and Elisa D. Mekler. 2023. How does HCI Understand Human Agency and Autonomy?. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3544548.3580651
- [11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative Research in Psychology 3, 2 (Jan. 2006), 77–101. doi:10.1191/1478088706qp063oa Publisher: Routledge.
- [12] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health 11, 4 (Aug. 2019), 589–597. doi:10.1080/2159676X.2019.1628806 Publisher: Routledge _eprint: https://doi.org/10.1080/2159676X.2019.1628806.
- [13] Zana Buçinca. 2024. Optimizing Decision-Maker's Intrinsic Motivation for Effective Human-AI Decision-Making. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, 1–5. doi:10.1145/3613905.3638179
- [14] Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2025. Contrastive Explanations That Anticipate Human Misconceptions Can Improve Human Decision-Making Skills. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. ACM, Yokohama Japan, 1–25. doi:10.1145/3706598.3713229

- [15] Federico Cabitza, Andrea Campagner, Chiara Natali, Enea Parimbelli, Luca Ronzio, and Matteo Cameli. 2023. Painting the Black Box White: Experimental Findings from Applying XAI to an ECG Reading Setting. Machine Learning and Knowledge Extraction 5, 1 (March 2023), 269–286. doi:10.3390/make5010017 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
- [16] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. 2022. BreastScreening-AI: Evaluating medical intelligent agents for human-AI interactions. Artificial Intelligence in Medicine 127 (May 2022), 102285. doi:10.1016/j.artmed.2022.102285
- [17] Tara Capel and Margot Brereton. 2023. What is Human-Centered about Human-Centered AI? A Map of the Research Landscape. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3544548.3580959
- [18] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23). Association for Computing Machinery, New York, NY, USA, 251–263. doi:10.1145/3581641.3584080
- [19] Christopher P. Cerasoli, Jessica M. Nicklin, and Alexander S. Nassrelgrgawi. 2016. Performance, incentives, and needs for autonomy, competence, and relatedness: a meta-analysis. Motivation and Emotion 40. 6 (Dec. 2016), 781–813. doi:10.1007/s11031-016-9578-2
- [20] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2024. Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. In Proceedings of the 29th International Conference on Intelligent User Interfaces. ACM, Greenville SC USA, 103–119. doi:10.1145/3640543. 3645199
- [21] Rebecca Crootof, Margot E. Kaminski, and W. Nicholson II Price. 2023. Humans in the Loop. Vanderbilt Law Review 76 (2023), 429. https://heinonline.org/HOL/Page?handle=hein.journals/vanlr76&id=447&div=&collection=
- [22] Mary L. Cummings. 2006. Automation and Accountability in Decision Support System Interface Design. *Journal of Technology Studies* 32, 1 (2006), 23–31. doi:10.21061/jots.v32i1.a.5
- [23] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–13. doi:10.1145/3544548.3580672
- [24] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In Equity and Access in Algorithms, Mechanisms, and Optimization. ACM, Boston MA USA, 1–23. doi:10.1145/3617694.3623261
- [25] Upol Ehsan, Koustuv Saha, Munmun De Choudhury, and Mark O. Riedl. 2023. Charting the Sociotechnical Gap in Explainable AI: A Framework to Address the Gap in XAI. Proceedings of the ACM on Human-Computer Interaction 7, CSCW1 (April 2023), 1–32. doi:10.1145/3579467
- [26] 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 Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22). Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3491101.3503727
- [27] Mica R. Endsley. 2023. Ironies of artificial intelligence. Ergonomics 66, 11 (Nov. 2023), 1656–1668. doi:10.1080/00140139.2023.2243404
- [28] Mica R. Endsley, Beth Bolte, and Debra G. Jones. 2003. Designing for Situation Awareness: An Approach to User-Centered Design. CRC Press, London. doi:10.1201/9780203485088
- [29] Lena Enqvist. 2023. 'Human oversight' in the EU artificial intelligence act: what, when and by whom? Law, Innovation and Technology 15, 2 (July 2023), 508-535. doi:10.1080/17579961.2023.2245683 Publisher: Routledge _eprint: https://doi.org/10.1080/17579961.2023.2245683.
- [30] Cedric Faas, Richard Bergs, Sarah Sterz, Markus Langer, and Anna Maria Feit. 2024. Give Me a Choice: The Consequences of Restricting Choices Through AI-Support for Perceived Autonomy, Motivational Variables, and Decision Performance. doi:10.48550/arXiv.2410.07728 arXiv:2410.07728 [cs].
- [31] Batya Friedman. 1996. Value-sensitive design. interactions 3, 6 (1996), 16-23. doi:10.1145/242485.242493
- [32] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In 27th International Conference on Intelligent User Interfaces. ACM, Helsinki Finland, 794–806. doi:10.1145/3490099.3511138
- [33] Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. 2025. Human-AI collaboration is not very collaborative yet: a taxonomy of interaction patterns in AI-assisted decision making from a systematic review. Frontiers in Computer Science 6 (Jan. 2025), 1–15. doi:10.3389/fcomp.2024.1521066 Publisher: Frontiers.
- [34] Google. 2025. Google PAIR. People + AI Guidebook. https://pair.withgoogle.com/guidebook
- [35] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review 45 (July 2022), 105681. doi:10.1016/j.clsr.2022.105681
- [36] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. Proceedings of the ACM on Human-Computer Interaction 3, CSCW (Nov. 2019), 1–25. doi:10.1145/3359280
- [37] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. Metrics for Explainable AI: Challenges and Prospects. doi:10.48550/arXiv.1812.04608 arXiv:1812.04608 [cs].
- [38] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (CHI '99). Association for Computing Machinery, New York, NY, USA, 159–166. doi:10.1145/302979.303030
- [39] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry* 11, 1 (Feb. 2021), 1–9. doi:10.1038/s41398-021-01224-x Publisher: Nature Publishing Group.

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

- [41] Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C.P. Snijders. 2023. It Seems Smart, but It Acts Stupid: Development of Trust in AI Advice in a Repeated Legal Decision-Making Task. In Proceedings of the 28th International Conference on Intelligent User Interfaces. ACM, Sydney NSW Australia. 528–539. doi:10.1145/3581641.3584058
- [42] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3581001
- [43] 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. ACM, Chicago IL USA, 1369–1385. doi:10.1145/3593013.3594087
- [44] Markus Langer, Kevin Baum, and Nadine Schlicker. 2024. Effective Human Oversight of AI-Based Systems: A Signal Detection Perspective on the Detection of Inaccurate and Unfair Outputs. Minds and Machines 35, 1 (Nov. 2024), 1. doi:10.1007/s11023-024-09701-0
- [45] Markus Langer, Veronika Lazar, and Kevin Baum. 2025. On the Complexities of Testing for Compliance with Human Oversight Requirements in AI Regulation. doi:10.48550/arXiv.2504.03300 arXiv:2504.03300 [cs].
- [46] 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. doi:10.1016/j.artint.2021.103473
- [47] 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 39 (Oct. 2023), 2853–2866. doi:10.1007/s00146-023-01777-z
- [48] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA. 1–15. doi:10.1145/3313831.3376590
- [49] Q. Vera Liao and Kush R. Varshney. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. doi:10.48550/arXiv.2110.10790 arXiv:2110.10790 [cs].
- [50] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3411764.3445260
- [51] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. Proceedings of the ACM on Human-Computer Interaction 5, CSCW2 (Oct. 2021), 1–45. doi:10.1145/3479552
- [52] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion 106 (June 2024), 102301. doi:10.1016/j.inffus.2024.102301
- [53] Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. 2025. Towards Human-AI Deliberation: Design and Evaluation of LLM-Empowered Deliberative AI for AI-Assisted Decision-Making. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, 1–23. doi:10.1145/3706598.3713423
- [54] Shuai Ma, Chenyi Zhang, Xinru Wang, Xiaojuan Ma, and Ming Yin. 2024. Beyond Recommender: An Exploratory Study of the Effects of Different AI Roles in AI-Assisted Decision Making. doi:10.48550/arXiv.2403.01791 arXiv:2403.01791 [cs].
- [55] Sara E. McBride, Wendy A. Rogers, and Arthur D. Fisk. 2011. Understanding the Effect of Workload on Automation Use for Younger and Older Adults. *Human Factors* 53, 6 (Dec. 2011), 672–686. doi:10.1177/0018720811421909 Publisher: SAGE Publications Inc.
- [56] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267 (Feb. 2019), 1–38. doi:10.1016/j.artint.2018.07.007
- [57] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI!: Hypothesis-driven Decision Support using Evaluative AI. In 2023 ACM Conference on Fairness Accountability and Transparency. ACM, Chicago IL USA, 333–342. doi:10.1145/3593013.3594001
- [58] Ozlem Ozmen Garibay, Brent Winslow, Salvatore Andolina, Margherita Antona, Anja Bodenschatz, Constantinos Coursaris, Gregory Falco, Stephen M. Fiore, Ivan Garibay, Keri Grieman, John C. Havens, Marina Jirotka, Hernisa Kacorri, Waldemar Karwowski, Joe Kider, Joseph Konstan, Sean Koon, Monica Lopez-Gonzalez, Iliana Maifeld-Carucci, Sean McGregor, Gavriel Salvendy, Ben Shneiderman, Constantine Stephanidis, Christina Strobel, Carolyn Ten Holter, and Wei Xu. 2023. Six Human-Centered Artificial Intelligence Grand Challenges. International Journal of Human-Computer Interaction 39, 3 (Feb. 2023), 391–437. doi:10.1080/10447318.2022.2153320 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2022.2153320.
- [59] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human Factors* 52, 3 (June 2010), 381–410. doi:10.1177/0018720810376055
- [60] Raja Parasuraman, Thomas B. Sheridan, and C.D. Wickens. 2000. A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 30, 3 (May 2000), 286–297. doi:10.1109/3468.844354
- [61] Sharon K. Parker. 2014. Beyond motivation: Job and work design for development, health, ambidexterity, and more. Annual Review of Psychology 65 (2014), 661–691. doi:10.1146/annurev-psych-010213-115208

- [62] Sharon K. Parker and Gudela Grote. 2020. Automation, Algorithms, and Beyond: Why Work Design Matters More Than Ever in a Digital World. Applied Psychology 69, 4 (2020), 956–1023. doi:10.1111/apps.12241
- [63] S. K. Parker and C. Knight. 2024. The SMART model of work design: A higher order structure to help see the wood from the trees. Human Resource Management 63, 2 (2024), 265–291. doi:10.1002/hrm.22200
- [64] Iryna Pentina, Tyler Hancock, and Tianling Xie. 2023. Exploring relationship development with social chatbots: A mixed-method study of replika. Computers in Human Behavior 140 (March 2023), 107600. doi:10.1016/j.chb.2022.107600
- [65] Muhammad Raees, Inge Meijerink, Ioanna Lykourentzou, Vassilis-Javed Khan, and Konstantinos Papangelis. 2024. From explainable to interactive AI: A literature review on current trends in human-AI interaction. *International Journal of Human-Computer Studies* 189 (Sept. 2024), 103301. doi:10.1016/j.ijhcs.2024.103301
- [66] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. Proc. ACM Hum.-Comput. Interact. 6, CSCW1 (April 2022), 83:1–83:22. doi:10.1145/3512930
- [67] Leon Reicherts, Zelun Tony Zhang, Elisabeth Von Oswald, Yuanting Liu, Yvonne Rogers, and Mariam Hassib. 2025. AI, Help Me Think—but for Myself: Assisting People in Complex Decision-Making by Providing Different Kinds of Cognitive Support. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. ACM, Yokohama Japan, 1–19. doi:10.1145/3706598.3713295
- [68] Richard M. Ryan and Edward L. Deci. 2000. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American Psychologist 55, 1 (2000), 68–78. doi:10.1037/0003-066X.55.1.68
- [69] Richard M. Ryan and Edward L. Deci. 2018. Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness. Guilford Publications. New York.
- [70] Thomas B. Sheridan. 2019. Individual Differences in Supervisory Control of Multiple Unmanned Vehicles. Human Factors 61, 4 (2019), 1–15. doi:10.1177/0018720819840003
- [71] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. International Journal of Human-Computer Interaction 36, 6 (2020), 495-504. doi:10.1080/10447318.2020.1741118
- [72] Ben Shneiderman. 2020. Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. International Journal of Human-Computer Interaction 36, 6 (April 2020), 495-504. doi:10.1080/10447318.2020.1741118 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2020.1741118.
- [73] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human-chatbot relationships. International Journal of Human-Computer Studies 168 (Dec. 2022), 102903. doi:10.1016/j.ijhcs.2022.102903
- [74] Peter J. Stanley, Nicola S. Schutte, and Wendy J. Phillips. 2021. A meta-analytic investigation of the relationship between basic psychological need satisfaction and affect. Journal of Positive School Psychology 5, 1 (April 2021), 1–16. doi:10.47602/jpsp.v5i1.210
- [75] 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 Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). Association for Computing Machinery, New York, NY, USA, 2495–2507. doi:10.1145/3630106.3659051
- [76] Minmin Tang, Dahua Wang, and Alain Guerrien. 2020. A systematic review and meta-analysis on basic psychological need satisfaction, motivation, and well-being in later life: Contributions of self-determination theory. PsyCh Journal 9, 1 (2020), 5–33. doi:10.1002/pchj.293 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pchj.293.
- [77] Anja Van den Broeck, D. Lance Ferris, Chu-Hsiang Chang, and Christopher C. Rosen. 2016. A Review of Self-Determination Theory's Basic Psychological Needs at Work. Journal of Management 42, 5 (July 2016), 1195–1229. doi:10.1177/0149206316632058 Publisher: SAGE Publications Inc.
- [78] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, Glasgow Scotland Uk, 1–15. doi:10.1145/3290605.3300831
- [79] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In 26th International Conference on Intelligent User Interfaces. ACM, College Station TX USA, 318–328. doi:10.1145/3397481.3450650
- [80] Katharina Weitz, Ruben Schlagowski, Elisabeth André, Maris Männiste, and Ceenu George. 2024. Explaining It Your Way Findings from a Co-Creative Design Workshop on Designing XAI Applications with AI End-Users from the Public Sector. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3613904.3642563
- [81] Wei Xu, Marvin J. Dainoff, Liezhong Ge, and Zaifeng Gao. 2023. Transitioning to Human Interaction with AI Systems: New Challenges and Opportunities for HCI Professionals to Enable Human-Centered AI. International Journal of Human-Computer Interaction 39, 3 (Feb. 2023), 494–518. doi:10.1080/10447318.2022.2041900 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2022.2041900.
- [82] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, Honolulu HI USA, 1–13. doi:10.1145/3313831.3376301
- [83] 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. ACM, Barcelona Spain, 295–305. doi:10.1145/3351095.3372852
- [84] Zelun Tony Zhang, Sebastian S. Feger, Lucas Dullenkopf, Rulu Liao, Lou Süsslin, Yuanting Liu, and Andreas Butz. 2024. Beyond Recommendations: From Backward to Forward AI Support of Pilots' Decision-Making Process. Proc. ACM Hum.-Comput. Interact. 8, CSCW2 (Nov. 2024), 485:1–485:32. doi:10.1145/3687024
- [85] Chaoyi Zhao and Wei Xu. 2025. Human-AI Interaction Design Standards. doi:10.48550/arXiv.2503.16472 arXiv:2503.16472 [cs].

A Example Task

Sample Solution

Question 1:

Consider this declaration of bar. For which values of x and y does it terminate? For which does it diverge? Assume the ideal interpreter (i.e., an interpreter that can work with arbitrarily large integers without overflowing).

1 let rec bar (x: int) (y: int) : int = if x = 5 then y else bar (x - 2) y

Answer 1:

bar terminates iff $x \ge 5$ and x is odd. For all even values of x and all x < 5, bar diverges. The value of y does not impact the termination of bar

(1 point for termination condition, 1 point for divergence condition)

Achievable Points: 2

Student Answer

Student ID: 9

Answer 1:

It terminates whenever $x \ge 5$ (because subtracting 2 repeatedly will eventually hit x = 5) and in that case returns y; it diverges whenever x < 5. The value of y is irrelevant (any integer y is returned upon termination).

Achieved Points: 0.0

B Prompts

We used the OpenAI o4-mini model to generate the 'student' answers for the 'Programming 1' exercises and to assign points to all student answers. To generate the students answers we prompted the OpenAI API with the following prompt:

"I have the following question for a test: *Question*. There are *Points* achievable points for this question. Generate an answer for this question, which a student would give that receives about *Percentage*% in a test. Please only provide the generated answer and no additional information."

For *Question* we inserted the question for the given exercise and for *Points* the amount of points one can achieve for this question. We prompted the OpenAI API 41 times for each question and varied the *Percentage* so that it aligned with the distribution of achieved grades in the psychology exam (21 x 100%, 15 x 75%, 5 x 50%).

For assigning points to the students' answers we used the following prompt structure:

"I have the following test question: *Question* and the matching sample solution: *Solution*. Achievable Points: *Points*. This is my answer to the question: *Answer*. How many points do I achieve with this answer to the question? Please only answer with the achieved points and do not provide any further information."

Here, we included the sample solution as *Solution* and the students answer as *Answer* additionally to the *Question* and *Points*.

C Instructions

C.1 Session 1

Purpose. Due to current advancements in the field of artificial intelligence, there are more and more systems that support humans in their decision-making. Technologies have been developed that can compete with human decision-makers in all kinds of decision-making tasks. These systems are already in use for court decisions, hiring decisions,

loan distribution, and more. Another domain where these technologies perform well is grading students' assignments, where they can support tutors and professors in their teaching. While there are AI systems that perform well in these tasks, there are concerns about using the AI's decisions without a human overseeing them. It is argued that humans are better at making decisions on edge cases or unusual situations, and also that humans employ ethical considerations in their decision-making. Based on such argumentation, many guidelines and laws require a human oversight person to oversee AI decision-making tools in high-stakes scenarios such as court decisions, loan distributions, and assessments of students' performance. In this workshop, we aim to identify the users' needs, desires, problems, and strategies when overseeing an autonomous decision-making system and prototype interfaces for supporting these. Performing such a task can be very tiring, frustrating, and hard. AI systems are only used if they are very good at their task, which makes it very difficult for human oversight people to detect any errors made by the system. Evidence shows that the argument supporting human oversight is flawed and that people are not able to increase the accuracy of the AI systems. We want to investigate which goals oversight people can have, which of them are achievable and motivating, and how the user interface can support the user to have realistic goals and how to achieve them.

Task. The workshop consists of two sessions with different goals. In the first session, you will be introduced to the task and have time to perform the task once. The task is to oversee an autonomous grading system that grades student tests. You will have 20 minutes to check if the autonomous system performed valid corrections or identify possible errors made by the system. After that, you will work in a group of people to reflect on and identify the specific strategies you used to approach the task and derive user needs and system requirements for an oversight interface. In the second session, you will brainstorm and select solutions for the needs and problems identified in the first session. After you have discussed different solutions, you will be provided with materials to create a paper mock-up of an interface that would support you in performing the grading oversight task. Both sessions will be moderated by an experimenter.

Collected Data. We will video and audio record both sessions and take pictures of the mock-ups and diagrams created by you. The recorded videos are not anonymized. They will be stored on the Perspicuous Computing Nextcloud, which is a local server running at Saarland University and will only be accessible to researchers involved in this project. Videos will be deleted upon completion of the research project or if you request deletion of your data. We might use pictures from the workshop in future publications, but will ensure that participants' faces are not recognizable. All other data will be pseudonymized. No explicit clues about your identity will be left in the stored data. The anonymized data will be released as part of the publications on this research project.

Duration and Breaks. Today's session will take about 2 hours. You will have 20 minutes to perform the oversight task at the beginning. Then we will create an affinity diagram together in a group discussion. You can take breaks at any time. If you are feeling tired, ask for a break. It is very important that you can always give your best! Of course you are free to end the experiment at any time.

Task Instructions. Imagine you are a Tutor of Programming 1 [Arbeits- und Organisationspsychologie]. Your job is to grade a test where 40 students participated, and which you will later hand out to the students. The test had four questions, and your Professor has given you the grading scheme. You are supported in this task by an automated grading system. It knows about the grading scheme and has assigned points to each student's answers. Due to the examination regulations of the university, a human oversight person is required to check if the autonomous system performed valid corrections or identify possible errors made by the system. Since the grading is already done by the automated system and you only need to check it, this should be faster than grading all the tests manually, and your Professor only

allocated 20 minutes for all the tests. In the next 20 minutes, you have time to familiarize yourself with the task. Keep in mind that it is not the goal to perform the task as good as possible in these 20 minutes. Rather, you should use the time to think about how you would approach it, whether your initial strategy will be efficient, and if it is enjoyable to use this strategy. You should also think about other strategies for this task, and how efficient and enjoyable they would be. Think about different goals you could have when performing this task and how achievable they are, how much you would enjoy performing the task with this goal in mind, and which strategies would help you to achieve the goal.

Affinity Diagram. Now we will create an affinity diagram to structure our thoughts and experiences. An affinity diagram is a visual tool used to categorize and structure large amounts of data or ideas. To create an affinity diagram, first, you will reflect on the strategies you took and the problems you faced when doing the task. You can write each thought on a sticky note. Try to use as few words as possible, so that everyone can read it from their seat. You will have enough time to explain your thoughts behind the keywords you wrote down. After collecting our thoughts individually, we will check each sticky note one by one and group them into categories, which are characterized by shared characteristics or relationships. Lastly, we will check for hierarchies and relations between the different categories.

Reflection Questions.

- What were you trying to achieve while checking the AI grading?
- What are other goals you could try to achieve when checking the AI grading?
- Walk us through the steps you would take to achieve each goal
- How did you feel while checking the AI grading and what made you feel this way?
- If you were telling a friend about checking the AI grading, how would you describe it?
- What would you do differently the next time you have to check the AI grading?
- In what ways is checking the AI grading different than grading the tests yourself?

C.2 Session 2

Purpose. In this session, we want to come up with ideas for a user interface that makes checking grades assigned by an AI easier and more fun. Remember the oversight task you performed in the last session. The goal of this session is to design a user interface prototype that would support you in performing this task in the future.

Task. Today, in the second session, you will brainstorm and select solutions for the needs and problems identified in the first session. After you have discussed different solutions, you will be provided with materials to create a paper mock-up of an interface that would support you in performing the grading oversight task.

Duration and Breaks. The study will take about 120 minutes. You will have 30 minutes to come up with solutions and decide on the solutions you want to integrate into your prototype. Then you have 50 minutes to work on a paper mock-up of an oversight interface. You can take breaks at any time. If you are feeling tired, ask for a break. It is very important that you can always give your best! Of course you are free to end the experiment at any time.

Crazy 8's. We will start this session with brainstorming solutions to make the oversight task from the previous session easier and more fun. We will use the crazy 8 method. This method is a brainstorming technique used to generate a large number of ideas in a short amount of time. It involves writing down 8 ideas within an 8-minute time frame,

without stopping to think about their feasibility or practicality. The goal is to produce a large quantity of ideas, which can then be evaluated and refined later.