Human Reliance on Decision Support Systems in High-Risk Scenarios

Marina Estévez Almenzar

TESI DOCTORAL UPF / 2025

Directors de la tesi
Ricardo Baeza-Yates and Carlos Castillo

Department and School of Engineering

Universitat Pompeu Fabra Barcelona

Dissertation submitted to the Department of Information and Communication Technologies of Universitat Pompeu Fabra in partial fulfillment of the requirements for the degree of:

DOCTORA PER LA UNIVERSITAT POMPEU FABRA

Social and Responsible Computing Group (https://www.upf.edu/web/recomputing), Department and School of Engineering, Universitat Pompeu Fabra, Barcelona, Spain.

Dr. Ricardo Baeza-Yates
(Thesis Supervisor)
Universitat Pompeu Fabra (UPF)
Barcelona, Spain

Dr. Carlos Castillo
(Thesis Supervisor)
Universitat Pompeu Fabra (UPF)
Barcelona, Spain

Thesis committee

Dr. Asia Biega
(Thesis Committee Member)
Max Planck Institute for Security and Privacy
Bochum, Germany

Dr. Daniel Gatica-Perez
(Thesis Committee Member)
Idiap Research Institute and EPFL
Martigny, Switzerland

Dr. Davinia Hernández-Leo
(Thesis Committee Member)
Universitat Pompeu Fabra
Barcelona, Spain

This doctoral project has been partially supported by: the Department of Research and Universities of the Government of Catalonia (SGR 00930); and the Maria de Maeztu Units of Excellence Programme CEX2021-001195-M, funded by MICIU/AEI/10.13039/501100011033.

Acknowledgments

Gracias,

A mi madre, por hacer que tu amor recorra kilómetros en solo unos segundos. A mi padre, por hacer que tu curiosidad sea un ejemplo constante cada día. A ambos: gracias por construir nuestro hogar y por tener el valor de animarnos a dejarlo cuando llegó el momento, aunque fuera doloroso.

A mi hermana, por tu incansable perseverancia, por superar todos los retos que te propones. A mi hermano, por tu bondad genuina, la más pura que he conocido jamás. A ambos: gracias por ser el mejor ejemplo que jamás tendré.

A Carlos Castillo y Ricardo Baeza, por guiar este trabajo.
A Emilia Gómez, por su calidez.

Mis primeros días en el departamento estuvieron marcados por las consecuencias de una estampida provocada por la primera ola de COVID en 2020. Quizás por eso cada persona que aparecía por la oficina parecía un nuevo tesoro en una isla desierta. Por orden (inexacto) de aparición:

A Olga, por compartir esta experiencia conmigo desde el primer día. A Francesco y Silvia, por hacer que nuestras risas aún resuenen en el 55.215. A Lorenzo Porcaro, por su sincero afecto y por llamarme "reina". A Adrià, por su gran corazón y su acento andaluz forzado. A Juan Gómez, por su honestidad y sus consejos siempre reparadores. A Aru, por nuestros almuerzos terapéuticos. A Tom, por nuestras charlas sinceras. A todas aquellas personas con las que pasé menos tiempo del que me hubiera gustado, y que aun así lograron mostrarme su cariño: Adri, Amelia, Laura, Pablo.

Aún con los ecos de la primera estampida, llegó la segunda: mucha gente se graduó. Pero las oficinas comenzaron a llenarse de nuevo con

gente maravillosa. En (también inexacto) orden de aparición:

A Roser, por ser mi apoyo y confidente, i per parlar-me en un preciós català del Maresme. A Sergio y Paula, por brindarme el aire curativo de las montañas. A Ioannis, por ser siempre el Best Captain. A Mykola, por compartir sus sabios consejos. A Abi, por los deliciosos dulces iraníes. A Anna Gatzioura y Jorge Saldivar, por ser los revisores más encantadores. A Nataly Buslón, por su tierna amabilidad. A Kalo y Alberto, por acompañarme en las dificultades de estos últimos meses.

Ha habido miedos y demonios en el camino, pero mis amigas han demostrado ser más fuertes que todos ellos. Me han enseñado cosas que son más importantes que cualquier otra cosa que pudiera aprender en mi recorrido académico.

A Ana y Rosa, por tener el don de reconocerme con solo unas pocas palabras, sin importar cuánto tiempo pase. A Noe y Javi, por ayudarme a desconectar en la pista de baile. A Juan, por quererme en tu vida. A Natalia, por llenarme el corazón cada vez que te veo. A Álida y Bárbara, por ser mi hogar. A Álida, porque si después de seis años viviendo juntas seguimos queriéndonos, es porque nos queremos mucho (aunque nos lo digamos poco). A Bárbara, por enseñarme las formas más bonitas del amor.

Abstract

This thesis investigates how human decision-making interacts with algorithmic decision support systems (DSS) in high-risk contexts as defined by the European Union's AI Act, focusing on facial recognition and criminal recidivism prediction. While AI systems are increasingly deployed in domains such as criminal justice, surveillance, and identity verification, the assumption that algorithmic recommendations can seamlessly and reliably augment human judgment lacks sufficient empirical grounding.

The dissertation presents a set of complementary empirical studies that examine the interplay between system accuracy, task difficulty, user familiarity, and targeted interventions. The first part analyzes overlaps and divergences between human and machine errors, revealing opportunities for hybrid human-AI decision-making strategies. The second part explores how DSS accuracy and task complexity affect performance in both face matching and recidivism prediction. Results show that accurate systems can improve decision quality, but low-accuracy systems exert disproportionate influence under difficult conditions, exposing users to automation bias. The third part investigates onboarding interventions, where brief exposure to DSS performance helps users calibrate trust and align their mental models with system capabilities. Findings demonstrate that onboarding enhances trust calibration and mitigates overreliance on poor systems.

Overall, this thesis contributes to the understanding of human-AI collaboration in high-risk scenarios, emphasizing that system design must account for human cognitive factors alongside technical performance. Rather than endorsing full automation, the findings advocate for hybrid decision-making frameworks that combine human judgment with algorithmic support to enhance effectiveness, fairness, and accountability while safeguarding against misuse and overreliance.

Resumen

Esta tesis investiga cómo la toma de decisiones humanas interactúa con los sistemas algorítmicos de apoyo a la toma de decisiones (DSS) en escenarios de alto riesgo, tal y como se definen en la Ley de IA de la Unión Europea, centrándose en dos casos de uso: el reconocimiento facial y la predicción de la reincidencia delictiva. Si bien los sistemas de IA se utilizan cada vez más en ámbitos como la justicia penal y la vigilancia, la suposición de que la IA puede complementar de forma fluida y fiable el juicio humano carece de una base empírica suficiente.

Esta tesis presenta un conjunto de estudios empíricos que examinan la interacción entre la precisión del sistema, la dificultad de la tarea, la familiaridad del usuario y las intervenciones específicas. La primera parte analiza las coincidencias y divergencias entre los errores humanos y los errores de máquina, revelando estrategias híbridas de toma de decisiones entre humanos e IA. La segunda parte explora cómo la precisión del sistema y la complejidad de la tarea afectan al rendimiento en ambos casos de uso. Los resultados muestran que los sistemas precisos pueden mejorar la calidad de las decisiones, pero los sistemas de baja precisión ejercen una influencia desproporcionada en condiciones difíciles. La tercera parte investiga las intervenciones de incorporación, en las que una breve exposición al sistema ayuda a los usuarios a calibrar la confianza y alinear sus modelos mentales. Los resultados demuestran que la fase de incorporación mejora dicha calibración y mitiga la dependencia desproporcionada.

Esta tesis contribuye a la comprensión de la colaboración humana-IA en escenarios de alto riesgo, haciendo hincapié en que el diseño de los sistemas debe tener en cuenta los factores cognitivos junto con el rendimiento técnico. En lugar de respaldar una automatización total, los resultados abogan por marcos híbridos de toma de decisiones que combinen el juicio humano con el apoyo algorítmico con el fin de prevenir el uso indebido y la dependencia excesiva.

Resum

Aquesta tesi investiga com la presa de decisions humanes interactua amb els sistemes algorítmics de suport a la presa de decisions en escenaris d'alt risc, tal com es defineixen en la Llei de IA de la Unió Europea, centrant-se en dos casos d'ús: el reconeixement facial i la predicció de la reincidència delictiva. Si bé els sistemes de IA s'utilitzen cada vegada més en àmbits com la justícia penal i la vigilància, la suposició que la IA pot complementar de manera fluida i fiable el judici humà manca d'una base empírica suficient.

Aquesta tesi presenta un conjunt d'estudis empírics que examinen la interacció entre la precisió del sistema, la dificultat de la tasca, la familiaritat de l'usuari i les intervencions específiques. La primera part analitza les coincidències i divergències entre els errors humans i els errors de màquina, revelant estratègies híbrides de presa de decisions entre humans i IA. La segona part explora com la precisió del sistema i la complexitat de la tasca afecten el rendiment en tots dos casos d'ús. Els resultats mostren que els sistemes precisos poden millorar la qualitat de les decisions, però els sistemes de baixa precisió exerceixen una influència desproporcionada en condicions difícils. La tercera part investiga les intervencions d'incorporació, en les quals una breu exposició al sistema ajuda els usuaris a calibrar la confiança i alinear els seus models mentals. Els resultats demostren que la fase d'incorporació millora aquest calibratge i mitiga la dependència desproporcionada.

Aquesta tesi contribueix a la comprensió de la col · laboració humana-IA en escenaris d'alt risc, posant l'accent que el disseny dels sistemes ha de tenir en compte els factors cognitius juntament amb el rendiment tècnic. En lloc de recolzar una automatització total, els resultats advoquen per marcs híbrids de presa de decisions que combinin el judici humà amb el suport algorítmic amb la finalitat de prevenir l'ús indegut i la dependència excessiva.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

As Artificial Intelligence[1] (AI) systems are increasingly embedded in high-risk[2] decision-making domains, concerns about safety, accountability, and trust have become central to both research and regulatory discourse. While AI excels in tasks requiring large-scale data processing and consistent execution, it often fails in areas where humans perform best: navigating ambiguity, contextualizing information, and exercising judgment under uncertainty. These contrasting capabilities have been revealed in recent research, suggesting that optimal performance can often be achieved not by choosing between human or AI, but by combining both through hybrid decision-making systems.

Yet, despite the theoretical advantages of such collaboration, true human-AI synergy remains rare and difficult to achieve in practice. The design of effective Decision Support Systems (DSSs) must account not only for technical performance, but also for the psycho-

[1]We understand *Artificial Intelligence* as defined in Article 3 of the AI Act (European Commission, 2024)

[2]By *high-risk* uses of AI, we refer to those categorised in Annex III of the AI Act (European Commission, 2024)

logical and behavioral factors that shape human interaction with the system. Studies using artificial or abstract tasks have helped identify key elements that influence trust and performance, such as the user's prior experience or the clarity of uncertainty communication. These factors significantly affect whether users appropriately rely on (or ignore) AI assistance.

Crucially, user reliance is not always aligned with system accuracy. Non-expert users, in particular, often exhibit mismatches between perceived and actual performance. Repeated interactions and early user experiences with the system have also been proved to play a defining role, sometimes locking users into patterns of overreliance or unwarranted skepticism that persist even as the system's performance changes.

In high-stakes domains the contextual complexity and potential consequences of AI errors demand greater scrutiny. Users are typically more cautious in these settings, often seeking human oversight and additional verification. Regulations like the AI Act by the European Commission (2024) reflects these concerns by imposing stricter obligations on high-risk applications. However, ensuring responsible hybrid decision-making in such contexts requires more than regulation; it demands collaboration paradigms that are sensitive to human behavior, cognitive load, and social responsibility.

Plenty of cognitive biases have been investigated in the literature related to hybrid decision-making. Users may demonstrate automation bias, blindly trusting AI suggestions. Others may show algorithmic aversion, reacting strongly to AI errors. Still others may selectively adhere to AI recommendations that confirm their prior beliefs. These biases complicate the task of fostering appropriately calibrated reliance and challenge the idea that transparency or explainability alone will lead to better outcomes.

Taken together, there is a critical need for AI systems that support (and not override) human decision-making. This means designing

DSSs that are not only technically robust but also psychologically and socially aware. It requires an understanding of how humans interact with algorithmic advice, how task characteristics influence reliance, and how interface design can nudge users toward better judgment. The path forward lies not simply in making machines technically stronger, but in making human-AI interactions more informed, adaptive, and ethically grounded.

1.2 Goals and Contributions

This thesis builds on these insights, exploring how human and machine errors differ, how task difficulty and user experience shape reliance on AI, and how targeted interventions can improve collaboration. The primary goal of this thesis is to advance the understanding and design of hybrid human-AI decision-making systems, with a specific focus on improving user interaction, trust calibration, and ethical robustness in high-risk domains. Motivated by the increasing deployment of AI in sensitive decision-making tasks and the corresponding regulatory challenges, this work investigates how human and algorithmic judgment can be meaningfully combined.

To this end, this dissertation makes the following key contributions, listed in Appendix B:

Conceptualization of Non-Human Errors
It introduces the notion of non-human errors – errors made by AI systems that are qualitatively distinct from human mistakes. This concept lays the foundation for better characterizing the complementary roles of humans and machines in decision-making contexts.

Empirical Evidence of Human-AI Complementarity
Based on a detailed analysis of face recognition tasks, the thesis shows that human and machine errors are often systematically different and can be leveraged to design more effective hybrid systems.

Behavioral Analysis of Human Reliance on AI
We investigate how users interact with AI-based Decision Support Systems (DSSs) in two high risk scenarios (face recognition and criminal recidivism prediction), under varying task difficulty and system reliability conditions. This reveals that users' misjudgment of system performance depends on perceived difficulty, making them vulnerable to automation bias.

Design of Trust Calibration Mechanisms
We evaluate the role of onboarding phases as tools to improve trust calibration in hybrid systems. We find that short initial exposure to the DSS and the task to be solved (especially when paired with ground-truth feedback) can enhance user discernment of system reliability.

Practical Implications for High-Risk Applications
By analyzing human behavior when interacting with a DSS in high-risk (*e.g.* criminal recidivism prediction), we offer actionable insights into the risks of over- and under-reliance on AI.

1.3 Organization of the Thesis

Figure 1.1 depicts the overall structure of this dissertation, summarised hereafter.

Chapter 2 reviews the literature on the field of human-computer interaction research and other related fields, with the aim of laying the groundwork for the research presented in the following chapters.

Chapter 3 of this dissertation introduces the concept of non-human errors: mistakes made by AI systems that are fundamentally different from those humans make, often violating human expectations and logic. These errors, although infrequent, can carry high risks, especially in sensitive applications like facial recognition and crimi-

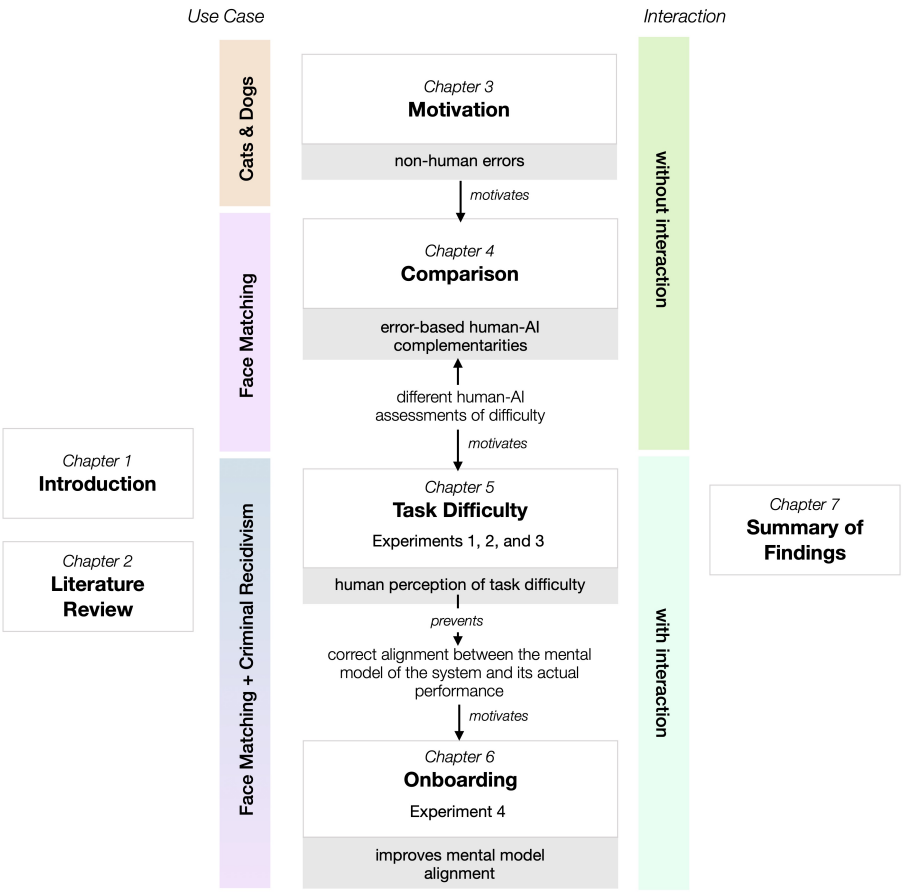


Figure 1.1: Conceptual scheme of the dissertation.

nal justice. The chapter presents a conceptual introduction to these errors, together with an empirical proof of concept that exemplifies the relevance of non-human errors.

Chapters 4, 5 and 6 include the empirical findings in this dissertation.

Building on Chapter 3, **Chapter 4** investigates how human and machine errors in face recognition tasks diverge, and how their complementarities can be systematically harnessed to design hybrid systems that are both more accurate and ethically sound. Our analysis reveals that human and algorithmic mistakes are neither redundant nor independent—they are meaningfully different, and these differences can be used to guide selective oversight. We show that minimal human intervention, strategically applied to the machine's most uncertain or error-prone cases, can significantly boost overall system performance. Yet we also highlight the influence of human biases, such as those linked to perceived gender or ethnicity, reinforcing the need for both technical rigor and ethical vigilance in hybrid decision-making systems.

Chapters 5 and 6 include the study of human behavior by exploring how humans interact with AI-based decision support systems (DSSs) in two complex, high-stakes tasks: face matching and criminal recidivism prediction. In **Chapter 5**, we examine how task difficulty and machine accuracy influence human reliance. Our results show that users struggle to recognize when a DSS is reliable, especially in cognitively demanding or unfamiliar tasks. We also observe that they are particularly vulnerable to automation bias when interacting with systems that begin with high accuracy but later degrade. These findings illustrate the urgent need to design AI systems not just for optimal performance, but for robust, user-aware interactions that help humans build appropriate levels of trust.

Chapter 6 addresses this challenge by evaluating the effectiveness of an onboarding phase (a brief period of initial contact with the task to be solved and the DSS) as a tool to calibrate user trust and improve hybrid decision quality. We find that onboarding enhances users' ability to discern system reliability, especially when it includes ground truth feedback. However, its effectiveness varies by task complexity and domain familiarity. Notably, in unfamiliar tasks like criminal recidivism prediction, onboarding can even influence unaided per-

formance over time, suggesting that interaction with AI may also function as a learning mechanism for human decision-makers when collaboratively solving unfamiliar tasks.

Finally, **Chapter 7** presents a summary of the findings, the several limitations of the research done and the open challenges and future perspectives.

The research plans that have enabled the experiments described in Chapters 4,5, and 6 were reviewed and approved by the Institutional Committee for Ethical Review of Projects (CIREP) at Universitat Pompeu Fabra. The review included compliance with internationally ethical principles and personal data protection guided by the EU General Data Protection Regulation (2016/679).

This manuscript also includes an appendix with complementary research to this work in **Appendix A** (Estévez Almenzar et al., 2022), carried out during my research stay at the Joint Research Centre in Ispra (Italy), and the contributions (list of publications and materials, presented in **Appendix B**).

2 Background and Literature Review

The integration of Artificial Intelligence (AI) into high-risk decision-making scenarios has transformed how complex decisions are made across diverse domains, from healthcare to finance and public policy. As AI technologies advance, understanding the entangled dynamics between human users and AI systems becomes increasingly critical for designing effective, trustworthy decision-making processes. In this chapter, we synthesize key research strands that explore the complementary strengths of humans and AI, the contextual factors shaping decision support in controlled and real-world tasks, and the behavioral mechanisms that influence how humans interact with algorithmic advice.

In §2.1, we examine the inherent complementarities between human cognitive capabilities and AI algorithms, highlighting how hybrid approaches can leverage the specific advantages of each to improve accuracy and robustness. In §2.2 we explore studies based on artificial or abstract decision-making tasks that shed light on fundamental factors affecting trust, reliance, and performance in human-AI collaboration, emphasizing the role of user expertise and interaction design. In §2.3 we consider the influence of task characteristics (risk level, expertise level, and complexity) on decision behaviors, illustrating how contextual demands shape user effectiveness and reliance on AI support.

Finally, in §2.4 this chapter addresses the critical role of human cognitive biases in mediating interactions with AI-driven Decision Support Systems (DSS). These biases, including automation bias, algorithmic appreciation and aversion, and confirmation bias, significantly impact users' acceptance, skepticism, and reliance on AI recommendations. Together, these themes provide a comprehensive foundation for understanding the challenges and opportunities inherent in human-AI decision-making systems and underscore the importance of designing DSSs that foster calibrated trust, transparency, and adaptive collaboration.

2.1 Human and AI Complementarities

Comparative studies between artificial intelligence (AI) models and human performance have become increasingly prominent across a range of domains, revealing capabilities and limitations on both sides. In the domain of visual perception tasks (such as interpreting visual data, or identifying and classifying objects), Dodge and Karam (2017a,b) demonstrated that algorithms perform comparably to humans on clean images, they exhibit a marked decline in accuracy when presented with distorted inputs. In contrast, human observers maintain high performance. In a similar vein, Phillips and O'Toole (2014) observed that algorithms outperformed humans in the case of static images, whereas humans demonstrated superiority in challenging static images and videos. This suggests that the human visual system is more robust, likely due to its use of global information. Subsequent research by Flachot and Gegenfurtner (2018) reinforced this view, revealing that humans outperform algorithms across a wide spectrum of image degradations. Moreover, as the level of distortion increases, the classification error patterns of humans and models diverge more significantly, indicating fundamental differences in perceptual processing.

Although advancements in training methodologies and dataset scaling have led to improvements in model performance, particularly in narrowing the gap with human-level accuracy, substantial discrepancies persist. For instance, a study by Geirhos et al. (2021) reported that despite progress, a considerable gap in image-level consistency remains, with humans and models exhibiting distinct error tendencies. Notably, much of the improvement observed was attributable to orders-of-magnitude increases in training data. Another study by Aust and Pons (2022) show that human performance in screen-based inspection tasks was superior to AI due to their strong cognitive abilities, decision-making capabilities, versatility and adaptability to changing conditions. However, in certain inspections that require high resource consumption, AI outperformed the human while being significantly slower. This suggests that the strength of AI systems lies in their consistency and availability.

Despite the comparative robustness of human perception in certain tasks, AI models have surpassed human performance in others. Lu et al. (2023) offer an example in distinguishing real from AI-generated images: human participants demonstrated a misclassification rate of 38%, whereas top-performing models exhibited significantly lower error rates at 13%, highlighting AI's superiority in specific detection tasks. Similarly, in the medical domain investigated by Tschandl et al. (2019), machine-learning classifiers have demonstrated superior diagnostic accuracy compared to human experts. Nonetheless, these models show reduced effectiveness when exposed to out-of-distribution data, pointing to a need for further research on model generalization. Hybrid approaches by Hekler et al. (2019) have shown promise in addressing these limitations. In one study, combining human and AI decision-making in classifying images suspicious of skin cancer resulted in a higher overall accuracy (82.95%) than either human judgment (42.94%) or AI alone (81.59%).

Beyond visual applications, comparisons have extended to educational contexts. Recent work by Banihashem et al. (2024) evaluating

feedback provided by ChatGPT and human peers on student essays revealed notable differences in content and focus. ChatGPT tended to offer more descriptive and form-based feedback, whereas peers emphasized problem identification within the essay content. These findings suggest a potentially complementary role for AI tools and human evaluators in educational assessment and feedback processes.

Collectively, these studies highlight that while AI models may surpass humans in tasks that benefit from scale and pattern recognition, human performance remains superior in contexts requiring robustness to variability, abstract reasoning, and contextual understanding.

Recent theoretical and empirical contributions have further highlighted the opportunities and challenges of combining human and machine intelligence. Korteling et al. (2021) emphasize that human cognition and AI are fundamentally different, shaped by evolution in the case of the former and engineered for computational efficiency in the latter. While humans struggle with tasks such as complex arithmetic, AI excels in such areas. This asymmetry suggests that the most effective use of AI lies not in attempting to replicate general human intelligence, but in developing systems that complement human cognitive limitations, particularly in areas such as logic, memory, and data analysis. Realizing this potential, however, requires what Korteling et al. (2021) call *intelligence awareness*: a clear understanding among professionals of AI's operational principles and its distinctness from human cognition, avoiding anthropocentric and anthropomorphic biases.

Empirical reviews of human-AI collaboration paint an even more complex picture. A large-scale analysis by Vaccaro et al. (2024) found that although AI can augment human performance (a phenomenon referred to as *human augmentation*), true synergy (in which human-AI teams reliably outperform both humans and AI alone) remains rare. Performance losses in collaborative settings are often due to overreliance (placing too much trust in AI systems) or underreliance (disregarding valuable AI input). Notably, the effectiveness of col-

laboration appears highly task-dependent: tasks involving content creation tend to benefit more from human-AI collaboration, whereas tasks requiring judgment or decision-making often result in decreased performance.

Taken together, these studies underscore both the remarkable progress of AI systems and the enduring strengths of human cognition. While AI models increasingly outperform humans in data-intensive or narrowly-defined tasks, human abilities remain essential for contextual reasoning, abstract judgment, and perceptual robustness. The emerging consensus suggests that hybrid approaches, designed with careful attention to cognitive asymmetries, collaboration dynamics, and task-specific demands, may offer the most promising path forward in advancing decision-making and performance across domains. This direction is a key motivation in this thesis.

2.2 Decision Support in Artificial Tasks

One of the most recognized human-AI hybrid perspectives in a decision-making context is a collaborative scenario, where the AI is commonly referred to as a Decision Support System (DSS). The interaction resulting from this collaboration manifests as strengths and weaknesses observed in the human and AI agents separately. Among the many experimental studies investigating human behavioral patterns when making decisions using a DSS, some involve generic or artificial tasks Lai et al. (2023). By *generic* or *artificial* tasks, we refer to those that either have no specific application domain or are fictitious tasks used to study research questions removed from a real-life context. These types of tasks provide a certain degree of neutrality that allows decision-making to be isolated from more complex aspects of reality, such as ethics or morals. The disadvantage is that the results obtained are difficult to generalize to more complex real-life contexts. Several fictional task studies have identified critical

variables that shape trust, reliance, mental model development, and overall task performance in hybrid decision-making contexts.

One prominent variable is user expertise (an aspect that we delve deeper in §2.3), which significantly influences preferences for DSS feedback and the alignment between perceived and actual performance. In a task involving estimation of news reading time with a DSS, Szymanski et al. (2021) found that experts and non-experts exhibited distinct preferences for the type and amount of feedback provided by the system. Notably, non-experts often demonstrated a mismatch between their actual performance when using the system and their subjective preference for it.

Another important variable lies in how response time and design decisions such as uncertainty visualization influence user trust, confidence, and decision accuracy. In a study involving the prediction of a water pipeline failure, Arshad et al. (2015) found that the way in which DSS uncertainty was visualized had a significant impact on user behavior. Specifically, different visualization methods altered levels of trust in the system, users' confidence in their own decisions, and ultimately the accuracy of those decisions.

Complementarily, another study by Park et al. (2019) examined how DSS response time affects user perception in a jellybean estimation task. The authors observed that shorter response times tended to boost user confidence when the system was accurate, while slower responses prompted more reflective thinking, especially when the system was inaccurate. These findings collectively highlight that subtle design choices in AI system behavior and interface can substantially shape the quality of human-DSS interaction.

Another set of studies by Yu et al. (2019) highlights the importance of time and interaction in fostering accurate mental models and appropriate levels of reliance in systems. In a glass-breaking prediction task, users were found to adjust their reliance dynamically based on observed system behavior within approximately 30 interaction tri-

als. Early interactions were especially crucial in shaping long-term influence, suggesting that the first phase of collaboration can heavily affects future reliance patterns.

Similarly, in a defect prediction task, Bansal et al. (2019a) observed that participants gradually refined their initially inaccurate or over-generalized mental models through repeated exposure to system feedback. Over time, users' decisions more closely approximated the DSS's true error boundaries, illustrating the critical role of iterative interaction and informative feedback in improving collaborative decision-making. However, collaboration is not static, and changes in the DSS itself (such as model updates) can introduce new challenges. In a follow-up study, Bansal et al. (2019b) observed that when the DSS was modified (e.g., via updated decision boundaries), even if the changes led to objectively improved performance, overall team performance could decrease. This suggests that temporal stability and predictability in DSS behavior may be more valuable to effective teamwork than marginal improvements in model accuracy.

The distribution of AI system errors and the sequence in which users experience system strengths and weaknesses have substantial implications for user reliance and cognitive alignment. In an activity recognition task designed by Nourani et al. (2021), users who initially observed the AI performing well developed inflated trust and exhibited automation bias, leading to more errors. Conversely, those who first encountered the system's weaknesses showed reduced over-reliance and improved judgment, indicating that early exposure to system limitations can foster more calibrated and effective collaboration.

2.3 Task Characteristics Influencing Decision Making

The dynamics of human-AI collaboration are profoundly shaped by the characteristics of the task itself. Three critical dimensions (risk or social consequences, user expertise or familiarity, and task difficulty or complexity) consistently emerge as key moderators of trust, reliance, and interaction quality in decision-making contexts. These factors influence not only how users perceive AI recommendations but also how they choose to engage with or defer to automated systems. Understanding the role of task characteristics is therefore essential for designing effective, context-sensitive AI decision support systems. Next, we examine how each of these dimensions impacts human behavior and system performance in hybrid decision-making settings.

2.3.1 Risk and Consequences

An important factor influencing human-AI decision-making dynamics is the level of risk associated with the task at hand. Research suggests that task risk can significantly mediate users' willingness to rely on automated support and their expectations for human oversight. Lai et al. (2023) emphasizes that findings derived from low-risk decision-making contexts, such as those derived from artificial tasks, may not generalize to high-risk situations, particularly in relation to user reliance on AI recommendations. This highlights the need for context-sensitive evaluations of DSS effectiveness across different risk profiles. The European Union's AI Act European Commission (2024) institutionalizes this insight through a risk-based regulatory framework, in which the use of AI systems are categorized according to their potential for harm. High-risk systems, such as those involved in critical infrastructure, education, or biometric identification, are subject to stricter governance and transparency requirements.

Complementing this regulatory approach, Green (2022) explores real-world applications in government settings, such as judicial sentencing and welfare assessments, where algorithmic decisions carry substantial societal consequences. In such domains, policymakers and scholars consistently advocate for robust human oversight to safeguard accountability and equity. The relevance of task-related social risk is further reinforced by Kim et al. (2023), which shows that when faced with high-stakes hypothetical scenarios, users tend to assess AI capabilities more critically and are more likely to seek additional verification, particularly when potential consequences (e.g., misidentification) carry significant social consequences.

Together, these studies illustrate that perceived task risk (whether technical, social, or institutional) substantially influences how humans engage with the DSS, underscoring the importance of aligning AI design and oversight mechanisms with the risk profiles of specific decision contexts.

2.3.2 Expertise or Familiarity

Another critical task-related factor influencing human-AI decision making is the user's level of expertise or familiarity with the task. Expertise often co-varies with task complexity and perceived risk, shaping users' interaction strategies and trust in AI systems. As noted in Lai et al. (2023), tasks vary significantly in the degree of domain expertise they require. While some tasks, such as artificial ones, can be performed with little to no prior knowledge, others, like cancer image classification, demand substantial training. These differences affect how users engage with AI, with novices and experts exhibiting both similar and distinct patterns of reliance and decision behavior.

Szymanski et al. (2025) proposes a conceptual framework that distinguishes stakeholder expertise from stakeholder role, highlighting

the need for decision support systems to account for this duality. By doing so, AI designers can better tailor interaction techniques to suit both technical and non-technical users across different contexts. Cheng et al. (2019) explores this further by showing that while interactive and transparent ("white-box") explanations can enhance non-expert comprehension, such interventions do not necessarily translate into higher trust. Interestingly, users' trust in algorithmic decisions is not affected by the explanation interface or their level of comprehension of the algorithm.

Complementing these findings, van Berkel et al. (2024) demonstrates that expertise influences user confidence more than actual accuracy in visual analytics tasks; notably, interaction techniques that reduce cognitive load were effective across both expert and non-expert groups. Schaffer et al. (2019) show that users' reported task familiarity predicts both increased trust and decreased reliance. Moreover, methods aiming to make the DSS understandable to the user, such as explanations, were only effective for users with low familiarity, while for more familiar users, these techniques introduced automation bias. Additionally, He et al. (2023) observes that participants prefer conversational AI assistants that dynamically adjust the level of support based on the user's evolving task familiarity. This preference for an adaptive assistance is rooted in the concept of building a mental model of the user, and supports more effective and human-centered AI interactions over time.

Collectively, these studies underscore that both actual expertise and subjective task familiarity significantly shape the cognitive, behavioral, and trust dynamics in human-AI decision making.

2.3.3 Difficulty or Complexity

Task difficulty and complexity are pivotal factors that shape how individuals interact with AI systems during decision-making processes.

These dimensions influence not only the degree of trust users place in the automated decision but also the extent to which they rely on these outputs and the interface modalities they prefer.

Empirical findings suggest that as task complexity increases, individuals tend to defer more to the DSS, even when such reliance may not be fully warranted. For instance, Bogert et al. (2021) found that in intellective tasks involving complex problem-solving, users demonstrate a greater preference for algorithmic advice over human consensus, especially under high difficulty conditions. This indicates a cognitive bias favoring algorithmic support when tasks surpass individual capability. This pattern of deference is echoed in operational environments such as baggage screening studied by Huegli et al. (2023). They observed that elevated task difficulty results in operators exhibiting excessive dependence on automated detection tools, often leading to omission errors.

The impact of task difficulty also extends to users' perceptions of system credibility and adaptability, as shown by Monroe and Vangsness (2022). In navigation tasks, increased cognitive load and stress reduce users' willingness to trust automated systems, particularly when those systems attempted to recover from their own errors. This suggests that difficulty may amplify user sensitivity to system fallibility. Moreover, Papenmeier et al. (2022) found that the perceived complexity of the task context can mediate how users interpret AI errors. When AI models failed on simple tasks, user trust deteriorated more than when failures occurred in difficult tasks, where users appeared more tolerant and attributed errors to the inherent complexity of the problem rather than system incompetence. However, Zhang et al. (2024) defend that increased complexity does not always enhance AI reliance in a productive way. The high mental workload associated with difficult tasks can drive overreliance, in which users surrender decision-making authority to the AI even when it may not be appropriate to do so. In contrast, low workload tasks elicit more balanced reliance patterns. Notably, the presence of explanations did

not mitigate overreliance in difficult cases, pointing to the limitations of explainability as a unique solution.

Gao et al. (2023) explores this further by observing discrepancies between users' perceived and actual reliance behaviors. In contexts such as qualitative coding supported by DSS, users reported perceiving greater helpfulness from the DSS for more difficult subtasks, yet paradoxically relied on the system more heavily in easier subtasks, potentially due to overconfidence in complex situations or underestimation of task demands. This disconnect underscores the importance of designing systems that account for both subjective perceptions and behavioral responses to difficulty.

The implications of task complexity are not limited to decision strategy but extend to interface design and user expectations. For instance, in business settings, He et al. (2023) found that users prefer simple conversational interfaces for routine, low-complexity tasks but favored more robust, graphical or multimodal interfaces when engaging with complex, information-dense problems. This suggests that effective human-AI interaction depends not only on the content and accuracy of recommendations but also on their delivery format, which should dynamically adapt to the task's cognitive demands.

Collectively, these findings highlight that task difficulty and complexity exert multifaceted effects on human-AI collaboration. They influence trust, reliance behavior, and user interface preferences. Designing AI systems that are sensitive to the evolving difficulty of the task at hand is therefore essential for optimizing human-AI joint performance, particularly in high-stakes or cognitively demanding domains.

2.3.4 Subjectivity, Moral, Hedonism

Another factor of the task that shape how individuals interact with AI systems during decision-making processes is the distinction between

objective and subjective tasks. Castelo et al. (2019) show that the more objective the task is perceived to be, the more likely the human is to be influenced by the machine. However, Logg et al. (2019) also find evidence that in some cases the distinction between objective and subjective tasks does not play an important role in how people are influenced by machine suggestions. In a similar vein, Mahmud et al. (2022) highlight the moral nature of the task at hand. People tend to move away from the machine when it comes to making moral decisions, such as those related to legal or medical issues (Bigman and Gray, 2018; Gogoll and Uhl, 2018; Bonnefon et al., 2024). Furthermore, it has been noted that when utilitarian results hold significant value, there is a preference for AI recommenders instead of human ones, whereas when hedonic aspects are prioritized, there tends to be a resistance to AI recommenders in favor of human decisions Longoni and Cian (2022).

2.4 Human Behaviors Influencing Decision Making

Cognitive biases play a central role in shaping how humans interact with a DSS. Despite growing advances in AI capabilities, human decision-makers remain susceptible to psychological tendencies that can distort judgment, influence trust, and affect reliance on algorithmic outputs. This section examines four key biases – automation bias, algorithmic appreciation, algorithmic aversion, and confirmation bias – each of which has been shown to affect decision quality in distinct but interconnected ways. Understanding these biases is essential for designing DSSs that promote calibrated trust and foster effective human-AI collaboration.

2.4.1 Automation Bias

Automation bias is defined as the tendency to overly rely on algorithmic outputs regardless of their correctness, and has emerged as a critical behavioral factor influencing decision quality in human-AI interaction. This cognitive bias can manifest as either commission errors (accepting incorrect AI advice) or omission errors (failing to act when AI fails to alert). The most recent literature presents a detailed view of the prevalence and impact of automation bias, shaped by user expertise, task context, and trust dynamics (Romeo and Conti, 2025).

However, some studies find limited evidence for automation bias in practice. For example, Alon-Barkat and Busuioc (2023) investigated decision-making in the public sector and found no general pattern of blind adherence to the DSS recommendations. The authors suggest that recent public controversies surrounding algorithmic discrimination may have fostered skepticism, reducing the influence of automation bias. However, the study did uncover a consistent pattern of what they call *selective adherence*: a related bias whereby participants accepted recommendations that aligned with pre-existing stereotypes, independent of whether the advice came from a human or an AI.

In the medical domain, evidence of automation bias is more robust and varies by expertise level. Gaube et al. (2021) found that even radiologists who rated AI-generated diagnostic advice as lower in quality still relied on it, illustrating a disconnect between perceived and actual reliance. Dratsch et al. (2023) and Kim et al. (2025) showed that less experienced clinicians were particularly vulnerable to commission errors when AI predictions were incorrect. This effect was less pronounced among highly experienced professionals, who maintained more stable diagnostic performance despite erroneous AI outputs.

Interestingly, overreliance on AI may not always stem from low expertise. Keding and Meissner (2021) examined senior executives in

high-stakes investment decisions and found that AI support increased confidence and decision quality, but also led to inflated perceptions of AI reliability, encouraging overreliance. This underscores how the perceived objectivity of AI systems can create a false sense of infallibility, even among experts.

Additional research highlights the individual differences that mediate automation bias. Küper et al. (2025) demonstrated that medical experience enhances self-reliance, while *dispositional* trust (a personality trait) can drive overreliance through its influence on *situational* trust. This aligns with findings from Duan et al. (2024), which showed that overreliance tends to be uniquely associated with AI systems, unlike human teammates. Users may inappropriately generalize trust from one AI interaction to others, reflecting a misapplied transfer of learned trust not typically observed in interpersonal trust dynamics.

2.4.2 Algorithmic Appreciation

A growing body of literature highlights the phenomenon of algorithmic appreciation, defined as the tendency of users to favor or trust algorithmic systems appropriately, particularly under certain cognitive or contextual conditions. This appreciation can enhance human-AI collaboration but also raises important considerations about users' critical engagement with algorithmic outputs, which could lead to automation bias. Algorithmic appreciation and automation bias are closely related and, based on this review of the literature, it appears that the terms are sometimes used interchangeably.

Wojcieszak et al. (2021) emphasizes the importance of contextualizing algorithmic appreciation within the broader sociotechnological landscape, arguing that as users become more familiar with AI systems through direct experience or increasing exposure, their trust in algorithms may grow, even in the absence of a clear understanding of their internal workings. This trust is not driven solely by

perceptions of AI objectivity, but also by cultural, experiential, and media-related influences, including portrayals of AI in popular fiction, as Sundar et al. (2016) observe. Such familiarity can, over time, reduce algorithm aversion and reinforce the notion of AI as a capable collaborator in decision-making processes.

However, the extent to which algorithmic appreciation consistently leads to adequate trust remains debated. For instance, Dikmen and Burns (2022) found that providing domain-specific knowledge to users enabled them to better interpret AI outputs and reduce dependence on incorrect predictions. However, this does not align with Poursabzi-Sangdeh et al. (2021), which explored the role of model interpretability in shaping user trust. Although participants who were shown clear, simplified models could better simulate the AI's behavior, they were paradoxically less able to detect and correct for major errors, likely due to cognitive overload or misplaced confidence in the model's transparency. These findings indicate that interpretability does not guarantee better human judgment.

Yang et al. (2020) further underscores the entangled relationship between trust, understanding, and decision quality. In their study, participants could not fully comprehend the underlying algorithmic mechanisms, yet well-designed visual explanations allowed them to make effective decisions with the AI's assistance. Trust and perceived understanding were positively correlated, suggesting that effective explanation design can facilitate appropriate algorithmic appreciation, even in the absence of deep technical understanding.

2.4.3 Algorithmic Aversion

While algorithmic systems are increasingly integrated into decision-making processes, their acceptance by human users is often hindered by a phenomenon known as algorithmic aversion: a tendency to distrust or reject algorithmic inputs. This behavior is shaped by a com-

plex interplay of cognitive, emotional, and contextual factors that significantly influence the effectiveness of human-AI collaboration.

One of the primary drivers of algorithmic aversion is the human reaction to AI errors. Jones-Jang and Park (2023) found that participants judged algorithmic mistakes more harshly than identical human mistakes, particularly when the AI violated expectations of flawless performance. This perceived unfairness was amplified in high-stakes contexts such as legal, medical, or hiring decisions. These findings suggest that users often hold AI systems to higher standards of perfection, and when these expectations are unmet, trust can quickly erode. Jones-Jang and Park (2023) argue that reducing unrealistic expectations (for instance, through transparent communication of the AI's limitations) may help mitigate such aversion.

The familiarity of the user with AI also influences aversion tendencies. Horowitz and Kahn (2024) showed that individuals with minimal understanding of AI were more prone to distrust it, while those with moderate understanding often overrelied on it, illustrating a Dunning-Kruger effect (Dunning, 2011). Only users with a high understanding of AI demonstrated balanced and calibrated reliance. These results highlight the need to promote AI education, while also encouraging critical thinking and calibrated expectations.

First impressions are another critical factor. According to Nourani et al. (2022), initial experiences with AI can shape long-term trust trajectories: positive early interactions increased tolerance for future errors, while early negative experiences led to lasting skepticism. This aligns with *ordering bias* literature and underscores the importance of ensuring robust early-stage AI performance in human-DSS applications.

Opportunities for user control and participation can also reduce aversion. Sharan and Romano (2020) and Dietvorst et al. (2018) reported that allowing users to make small modifications to algorithms enhanced trust, as it conveyed a sense of agency. Additionally, Dietvorst

et al. (2018) show that even when users observed AI systems outperform humans, they still exhibited aversion after witnessing a single AI error. These findings are reinforced by Dietvorst et al. (2015) and Bogert et al. (2021), who concluded that algorithmic decisions are more harshly punished than human ones when they fail.

A recurring theme across studies is the perception that AI fails to account for qualitative and contextual factors in decision-making. Langer and Landers (2021) emphasized that users believe algorithms overlook human-specific values and situational subtleties, contributing to a sense of decision incongruity. When people's opinions and contextual knowledge were integrated into the system, users expressed greater confidence and willingness to follow algorithmic recommendations (Kawaguchi, 2021). This suggests that participatory design may serve as a valuable strategy to mitigate aversion.

Another important determinant is the perceived role of the algorithm in the decision-making hierarchy. Zhang et al. (2021) found that users were more receptive to algorithmic input when it was framed as a support tool for human decision-makers rather than a replacement. Users preferred systems where final decisions remained in human hands.

Interestingly, even aesthetic and embodiment factors can trigger aversion. Laakasuo et al. (2021) demonstrated that people were less accepting of decisions delivered by a system with an obviously robotic design compared to one with a human-like appearance.

Emotional factors also play a substantial role. Zhang et al. (2021) introduced the concept of *emotional trust*, which involves feelings of comfort, security, and reassurance in the decision-making process. When such trust was lacking, users showed greater resistance to algorithmic advice. Relatedly, Prahl and Van Swol (2021) found that users experienced more negative emotions and assigned more blame to AI advisors than to human ones, especially when errors occurred. This emotional disconnect stems from a fundamental belief that machines are not meant to make subjective decisions.

Finally, Dietvorst and Bharti (2020) explored the variance bias in users' perceptions: people tend to believe that humans are more capable of achieving perfect or near-perfect decisions due to their greater variability, even though algorithms may perform more consistently on average. This belief feeds into the narrative that human decisions, despite being less reliable overall, hold more potential for excellence, leading to an undervaluation of algorithmic reliability.

2.4.4 Confirmation Bias

Confirmation bias, the tendency to favor information that confirms one's pre-existing beliefs or hypotheses, has been identified as a significant factor in shaping how individuals interact with DSSs. Confirmation bias can influence both the interpretation and acceptance of algorithmic advice, often reinforcing rather than challenging users' initial judgments.

Gaube et al. (2021) investigated how diagnostic advice from AI versus human sources influenced both perception of advice quality and diagnostic accuracy among physicians of varying expertise levels. The authors found a general tendency among both expert radiologists and less experienced physicians rely on initial judgments when processing subsequent advice. This pattern was particularly pronounced among less experienced professionals, who showed greater susceptibility to anchoring effects and confirmation bias regardless of whether the recommendation came from an algorithm or a human.

In support of this view, two studies explicitly found no direct evidence of automation bias, emphasizing instead the dominant role of confirmation bias in shaping decision-making (Selten et al., 2023; Bashkirova and Krpan, 2024). Selten et al. (2023) examined how AI recommendations influenced the behavior of police officers, and found that they were significantly more likely to trust and follow AI suggestions when they were congruent with their professional intu-

ition.

Similarly, Bashkirova and Krpan (2024) explored the integration of AI in diagnostic processes within mental healthcare. Their findings mirrored those of Selten et al. (2023): participants demonstrated a clear preference for AI recommendations that confirmed their initial clinical assessments. Notably, practitioners who perceived themselves as highly experienced exhibited greater skepticism when AI outputs diverged from their professional evaluations.

Building on the existing literature, this dissertation addresses several key gaps at the intersection of human cognition and algorithmic decision support, particularly in high-risk domains identified in the AI Act by the European Commission (2024), such as facial recognition and criminal recidivism prediction. While prior studies emphasize the importance of task characteristics, user expertise, and cognitive biases in shaping human-AI interaction, a considerable part of this work has relied on artificial tasks or abstract settings. This thesis extends the field by investigating these dynamics in socially consequential, high-stakes contexts. It also offers a finer-grained analysis of how system accuracy, task difficulty, and user familiarity interact to influence human reliance, joint performance, and human susceptibility to automation bias – variables that have been explored individually in the literature but rarely examined in combination or under real-world constraints.

A novel contribution of this work lies in its experimental focus on interventions that can foster calibrated trust and improve human-AI collaboration, such as brief, feedback-rich onboarding phases. While previous research has highlighted the role of repeated interaction in shaping trust, this dissertation shows that even minimal exposure to system behavior – especially when ground truth is available – can help users form more accurate mental models of a DSS. In addition, this work shows evidence of what we refers to as *implicit learning* through sustained interaction with input systems in unfamiliar tasks,

a phenomenon that, to our knowledge, has been overlooked in previous research. By showing that onboarding can both mitigate overreliance and intervene user learning in complex scenarios, this research advances the design of decision support systems that are not only transparent and adaptive but also capable of reinforcing responsible human oversight.

3 Defining Non-Human Errors

This chapter is based on Baeza-Yates and Estévez-Almenzar (2022)

3.1 Introduction

Imagine that you enter a skyscraper and the elevator has a sign that says: "Works 99% of the time". Would you take the elevator? Arguably, most people would not. However, if the sign says "Does not work 1% of the time and when that happens, stops", more people probably would use it, because of the possibility to evaluate the consequences. Today, Artificial Intelligence (AI) models are mostly evaluated on the basis of success rather than failure. Worse, this evaluation rarely takes into account the potential harm of its mistakes – a common practice in the pharmaceutical and food industries, among others.

Along the same lines as this example, the reliability of recent benchmarks in accurately reflecting human-centered tasks has been questioned (Tsipras et al., 2018, 2020; Bender et al., 2021). The practice of focusing model evaluation on accuracy has been described as a dangerous habit (Northcutt et al., 2021; Geirhos et al., 2020) that

ignores some important aspects of human perception when developing a solution to a problem, such as carefully studying the risks of the solution and its different points of operation. This lack of diversity in evaluation procedures reinforces the difference between human and machine perception of the relevance of data features (Ilyas et al., 2019).

The benchmark-task misalignment can be explained by the misalignment between human and machine perceptual mechanisms. In this preliminary study we propose a simple error-based taxonomy to bridge those differences that could be potentially harmful, so to achieve more reliable model training and evaluation procedures, even if it implies a decrease of the accuracy.

To do that, it is essential to drive a decentralization of the evaluation process in AI models. To do that, we propose to study the different types of errors that an AI model can make, focusing on how they differ from those errors that a human might make.

3.2 Non-Human Errors

To illustrate in a simplified way the error exploration that we are proposing, let us consider a problem with a binary solution space. In this space, we are given the ground truth, so we are capable of determining whether a point from the space is a correct or an incorrect answer.

In Figure 3.1 we can see this ground truth represented as a green sphere, positioned on the solution space, such that those points that fall into the green area are the true positive answers, and the rest of the points are the true negative answers. We can see two more shapes in this space that symbolize the perceptual agents; a red triangle representing the machine predictions, and a blue ellipse representing the human predictions. Following the previous logic, in Figure 3.1b we

(a) (b)

Figure 3.1: Visual representation of human and AI performances. On the left, the red triangle represents the ML model, the blue ellipse represents the human, and the green circle represents the ground truth. They are positioned in the solution space of a binary prediction problem. For every figure, *positive* answers are inside, and *negative* answers are outside, with the correct answers determined by the green circle. On the right we can see the yellow region representing the correct answers obtained by both the human and the AI model.

can see the true answers correctly predicted by both the model and a human. In Figure 3.2, we focus on the errors. Here we can distinguish between two kinds of errors: false positives and false negatives. And we make another distinction based on the entity that is making the error (human and/or AI model).

In this abstract scenario, where no concrete use case is specified, we wonder whether we can determine which errors are the most relevant in terms of human harm risk. It is reasonable to think that the errors related to harmful consequences for humans are those that are unexpected and atypical for them. These errors are difficult for us to explain and control. Since, as humans, we are accustomed to human errors, we might expect that those errors that are furthest away from the errors that a human might make could be considered risky: we refer to these types of disparate errors as *non-human* errors (see Figure 3.3).

(a) (b) (c) (d) (e) (f)

Figure 3.2: Visual representation of false negatives and false positives attributed to the model, the human, or to both. The first three diagrams – (a), (b) and (c) – represent in yellow false negatives errors done by only the model, only the human, or by both, respectively. The next three diagrams – (d), (e) and (f) – represent in yellow the false positives errors done by only the model, only the human, or by both, respectively. Error areas are overstated to emphasize the idea.

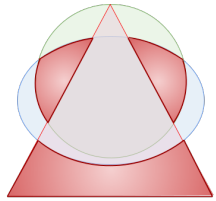


Figure 3.3: Non-human errors stressed in red: both false negatives and false positives done by the AI model but not by humans – cases (a) and (d) in Figure 3.2).

We can also formalise this idea in terms of mathematical sets. This will help us to formally define the different types of errors mentioned and graphically expressed above. We denote S as the green sphere, T as the red triangle, and E as the blue ellipse (see in Figure 3.1a). Following the logic explained above, we could consider these sets of points in the solution space (and their complementary sets, noted as

$\overline{S}$, $\overline{T}$, and $\overline{E}$ respectively) as follows:

$$
\begin{aligned}
S &\equiv \text{true positives} \\
T &\equiv \text{positives predicted by the model} \\
E &\equiv \text{positives predicted by the human} \\
\overline{S} &\equiv \text{true negatives} \\
\overline{T} &\equiv \text{negatives predicted by the machine} \\
\overline{E} &\equiv \text{negatives predicted by the human}
\end{aligned}
$$

Focusing on the errors shown in Figure 3.2, now we denote the false positives errors made only by the machine as P_m (Figure 3.2d), and we define this set of errors as

$$P_m = T \cap \overline{E} \cap \overline{S}$$

Similarly, we denote and define the false positives errors made only by the human (Figure 3.2e), the false positive errors made by both the machine and the human (Figure 3.2f), the false negative errors made only by the machine (Figure 3.2d), the false negative errors made only by the human (Figure 3.2b), and the false negative errors made by both the machine and the human (Figure 3.2c) as follows, respectively:

$$
\begin{aligned}
P_h &= \overline{T} \cap E \cap \overline{S} \\
P_b &= T \cap E \cap \overline{S} \\
N_m &= \overline{T} \cap E \cap S \\
N_h &= T \cap \overline{E} \cap S \\
N_b &= \overline{T} \cap \overline{E} \cap S
\end{aligned}
$$

Note that all these sets are disjoint because of the exclusivity imposed when considering which agent commits the error. Now the sets of interest arise from the union of some of the previous sets. We note

M as the non-human errors (those committed by the machine but not by the human) explained above, H as those errors committed by the human but not by the machine, and B as those errors committed by both the human and the machine together:

$$M = N_m \cup P_m$$
$$H = N_h \cup P_h$$
$$B = N_b \cup P_b$$

Here we focus on M, non-human errors, which we believe are the errors that should be addressed first because of the harm risks that could be involved in making errors that escape human logic. But how can we precisely determine these errors? How can we measure how far an answer should be from human logic in order to call it a non-human error? We address these challenges in the next section, which also serves as proof of concept for non-human errors.

3.3 Proof of Concept

Approaching a problem by adopting the previous abstract perspective allows us to visualize it with some independence from the use case or real-world application, which is good for understanding the wide range of operating points. The distinction of non-human errors is based on the distinction between the successes and mistakes made by the different perceptual agents (human and machine). The category of non-human errors can be found in those human centered tasks that can be at least partially solved in a reasonable way by the humans and where an AI algorithm is applied instead. However, in practice, consideration of the specific use case will be decisive.

We next apply this idea to a simple but illustrative problem: classifying images of dogs and cats according to their breed (Parkhi et al.,

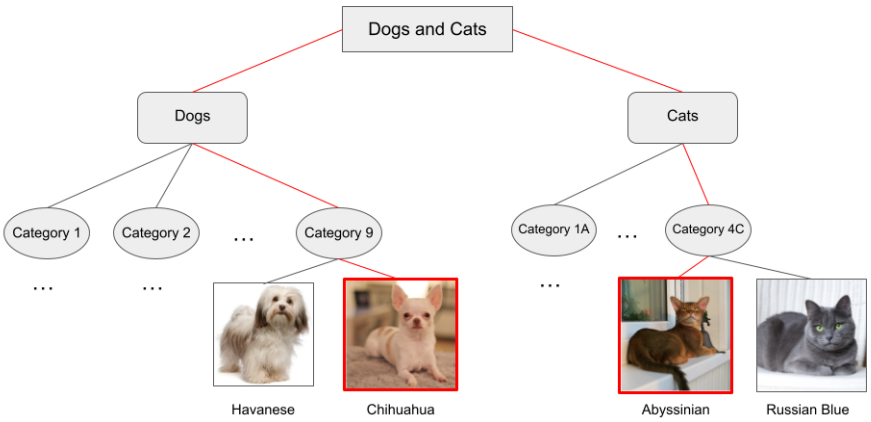
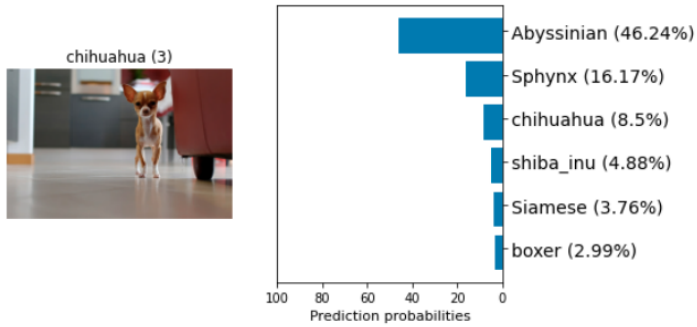


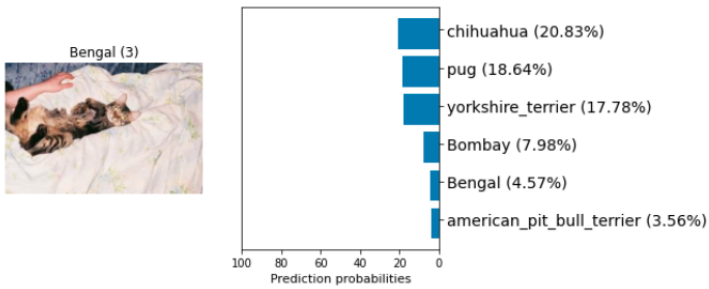
Figure 3.4: Part of the error taxonomy obtained from the analysis made with Oxford-IIIT Pets data set (Parkhi et al., 2012). In red, one of the most common non-human errors committed by the BiT model (Kolesnikov et al., 2020).

2012). This translates into a fine-grained image classification problem that can be addressed by using deep neural networks. Based on expert sources in the classification of these animals (FIFe and FCI Federations), we construct a taxonomy that represents the possible errors that can be made in this task. Following our definition of non-human errors, in this problem we can identify them as those errors that are fundamentally different from the errors that a human solving this task would commit. Therefore, we define as non-human errors those cases in which the machine classifies a dog as a cat, or vice versa (see Figure 3.4). Notice that there might be other non-human errors when comparing among only cats or dogs, but those are much less important and less common than the definition that we use for this proof of concept and provides a lower bound for non-human errors.

We have selected one of the top-ranked algorithms for solving this



(a)



(b)

Figure 3.5: Two of the non-human errors obtained when running the BiT model (Kolesnikov et al., 2020) over the Oxford-IIIT Pets data set (Parkhi et al., 2012). In (a) a Chihuahua is mistaken for an Abyssinian cat with a confidence of 46.24%. In (b) a Bengal cat is mistaken for a Chihuahua with a confidence of 20.83%, a percentage very close to the one of the second option in the list of breeds sorted by their probability of being selected as the tag for that image.

specific task, the Big Transfer (BiT) model from Kolesnikov et al. (2020), which achieved 93% of accuracy. In Figure 3.6 we give the full confusion matrix of 3,312 prediction pairs among 25 dog breeds (top-left) and 12 cat breeds (bottom-right), where we can see that there are 4 pairs that are hard to classify (two breeds of Terriers and

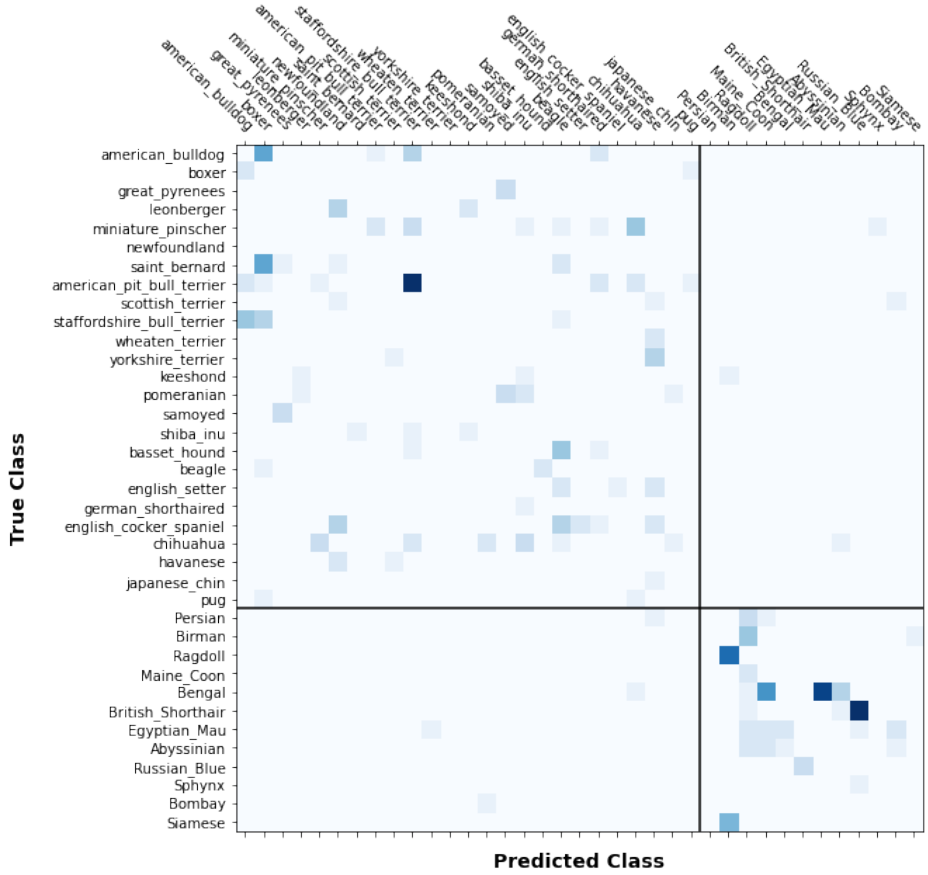


Figure 3.6: Confusion matrix for the Oxford-IIIT Pets data set (Parkhi et al., 2012). Darker the squares, more errors were made for that pair of breeds.

3 pairs of cat breeds). Notice that this confusion matrix is in general non-symmetric, as the output of the model may differ because the input and the prediction for each pair is different.

Here we found that more than 3% of the errors were non-human errors (8 of 241 errors), which appear as light squares in the top-right and

bottom-left of Figure 3.6. Two of these errors are shown in Figure 3.5, where a Chihuahua is classified as an Abyssinian cat (Figure 3.5a) and a Bengal cat is classified as a Chihuahua (Figure 3.5b). However, there is a notable difference between these two errors: the certainty of the answer provided by the algorithm.

3.4 Discussion

The results of our proof of concept illustrate a key limitation in conventional AI evaluation: high overall accuracy does not guarantee reliability in human-centered contexts. Despite achieving 93% accuracy on a fine-grained classification task, the model still produced errors that a human expert would be unlikely to make, such as classifying a dog as a cat or vice versa. While these misclassifications constituted a relatively small percentage of total errors (just over 3%), their qualitative impact may be disproportionately high, particularly in applications where trust and safety are critical.

This finding reinforces the claim that accuracy alone is insufficient as a proxy for model quality, especially in domains that intersect with human judgment and expectations. From a human-centered design perspective, non-human errors can undermine the usability and perceived trustworthiness of AI systems. In our taxonomy, these errors represent not just misclassifications, but perceptual divergences in how humans and machines interpret the same input.

It is also worth noting that our framework does not assume humans are error-free or superior in all tasks. Rather, it assumes that in human-centered applications, errors that resemble human reasoning failures are more comprehensible and manageable. In contrast, those errors lying outside the bounds of typical human performance represent a potential source of risk precisely because they are unexpected and difficult to predict or correct.

While our proof of concept focused on image classification, the framework can be extended to a range of tasks, including high-stakes scenarios. One concrete such example happened in 2018, when a Uber self-driving car was not able to recognize a woman in a bicycle crossing a road at night in Tempe, Arizona.[1] A human most probably would have recognized the woman and hence this is a non-human error. We do not know if the backup driver could have reacted on time, but she was seeing a video as the car was working well until then. Finally, she was charged of negligence, as Uber quickly settled with the family of the victim to avoid being sued (Smiley, 2022).

In the remainder of this dissertation, we will work with two use cases categorized as high risk in the recent AI Act by the European Commission (2024): facial recognition (Chapters 4 to 6) and criminal recidivism prediction (Chapters 5 and 6). The difference between human errors and machine errors mentioned in this chapter, as well as the existence of non-human errors and their potential harm, motivate the following comparative studies between human errors and machine errors, carried out with facial recognition tasks.

[1] *Self-driving Uber kills Arizona woman in first fatal crash involving pedestrian*, The Guardian, 19 March 2018.

4 An Error-Based Study of Human-Machine Complementarity

This chapter is based on Estévez-Almenzar, M., Baeza-Yates, R., and Castillo, C. (2025). *A Comparison of Human and Machine Learning Errors in Face Recognition* (TRUST-AI @ ECAI 2025)

4.1 Introduction

Decision support systems (DSS) powered by machine learning (ML) are increasingly used in high-stakes scenarios including immigration Chun and Wai (2007); Akhmetova and Harris (2021), video-surveillance Fontes et al. (2022); Park and Jones-Jang (2023), healthcare Prentzas et al. (2023), justice Green (2022), and access to labor and education Frank et al. (2019); Zhai et al. (2021), among many others. In these application domains, ML systems should not be autonomous, but rely on a human operator or expert, who should be responsible for the final decision. An in-depth understanding of the dynamics of human-algorithm interactions is crucial for developing safe, trustworthy systems (Matias, 2023). A crucial step before designing the interaction paradigm is understanding the complementarities between human and machine intelligence. In an "ideal" sce-

nario, there would be perfect complementarity: cases challenging for the ML system are easily handled by the human operator, and vice versa. Conversely, the worst case is when there is total overlap: cases that are difficult or uncertain for the ML also lead to human errors. In practice, we may find applications that are somewhere between these extremes, as our empirical findings demonstrate within the context of face recognition.

Motivated by the ideas and preliminary results introduced in the previous chapter, we observe that little has been done to understand when human and machine errors are similar and when they are different. As studied by Jussupow et al. (2020); Dietvorst et al. (2015), analyzing these similarities and differences is particularly important because the presence of algorithmic errors influences well-known patterns of human-machine interaction, such as *algorithmic aversion*, a biased and overly negative human evaluation of an algorithm. The question arises as to whether this aversion also varies depending on whether or not the errors presented by the model resemble those that a human agent might make. If we are able to avoid this type of bias, there is another risk: *automation bias,* an over-reliance on automated decision support mechanisms, widely studied in the literature (Lyell and Coiera, 2017; Cummings, 2017; Dratsch et al., 2023; Kim et al., 2025).

Our goal in this chapter is to compare ML errors and human errors when solving a specific task: face matching. This task allows us to transfer the central idea of the previous chapter – studying the differences and similarities between human and machine errors – to a real, high-risk context. We hypothesize that the results of this analysis will allow us to propose an oversight strategy that optimally combines the strengths and weaknesses of the ML system and the human.

The use of the concept "human error" could suggest an homogeneity that is almost non-existent in real life. Human perception varies from individual to individual, either due to variations in physiologi-

cal structures or external influences such as culture (Davidoff et al., 2008). It becomes even more complex when it comes to assessing human perception in distinguishing between other human identities, such as in the context of a face recognition task. We tackle this complexity through a demographically diverse user study for face matching in which possible inter-individual differences and disagreements are considered. Proposing hybrid human-machine strategies in the field of face recognition is crucial. For instance, in an automated system integrated in a police surveillance scenario, errors deserve special attention. In January 2020, Robert Julian-Borchak Williams became the first documented person in the U.S. that was wrongfully arrested based on a false hit produced by facial recognition technology.[1] Many more cases have been documented, and often there are racial biases.[2]

Our main findings for the face recognition task we study are: (1) humans rarely produce false positives; (2) the ML similarity score is a potential error predictor; (3) humans find it easier to address mistakes made by an individual model compared to addressing shared errors between two models; and (4) in face recognition the human perception of gender expression and ethnic appearance is determinant. These findings provide a method for detecting potential errors in automated facial recognition, and help us find those errors that a human annotator has a high chance of correcting. Applying this approach in a practical setting enables us to develop an effective evaluation strategy that maximizes joint human-machine accuracy while controlling human annotation effort. Unlike other approaches that strictly emphasize the enhancement of accuracy through algorithmic advancements, this work underscores not only the importance of incorporating the human factor in the race for accuracy maximization, but also the effectiveness of this approach.

The rest of the paper is organized as follows. In §4.2 we review related work, followed by our research questions and our methodology in §4.3

[1] *Wrongfully Accused by an Algorithm*, The New York Times, 24 June 2020.
[2] *When AI Gets It Wrong.* Innocence Project, 19 September 2023.

and §4.4, respectively. In §4.5 we present our results, while in §4.6 we discuss our results, limitations, and give our conclusions §4.7.

4.2 Related Work

In addition to the reviewed literature presented in Chapter 2, here we review further research related to this chapter.

Human and ML performance ML systems may outperform human annotators in tasks considered simple or moderately difficult, but they tend to struggle when faced with more complex conditions that mirror real-life scenarios. White et al. (2015) observed that in challenging tasks, non-expert observers showed performance comparable to that of some facial recognition algorithms, while in some cases experts outperformed these algorithms. In a similar vein, Rice et al. (2013) observed that some systems did not make accurate identifications, while humans exceeded random chance. Throughout the years, numerous competitions have been conducted to assess human-algorithm performance in different face recognition tasks, opposing algorithmic accuracy to human accuracy, thus establishing a distinction between two solving agents that, instead of collaborating, compete. Phillips and O'Toole (2014) conducted a cross-modal study to evaluate the results of selected human-algorithm competitions in facial recognition. Their findings revealed that algorithms outperformed humans in the case of simple frontal static images, whereas humans demonstrated superiority in challenging static images and videos. Rice et al. (2013) focused on particular scenarios where facial recognition systems were not successful and examined human performance under these circumstances. They found cases where facial recognition algorithms failed, yet humans performed better than random guessing. White et al. (2015) observed that in forensic facial

recognition algorithms performed similarly to certain observers and were outperformed by experts.

Combining human and machine intelligence Researchers have seek to uncover how human and machine intelligence can be reliably combined. *Algorithm aversion* has been extensively studied (Dzindolet et al., 2002; Jussupow et al., 2020; Reich et al., 2023; Jones-Jang and Park, 2023; Horowitz and Kahn, 2024), and has been shown by Dietvorst et al. (2015) that it becomes particularly noticeable when users witness mistakes made by the algorithm. As Dietvorst et al. (2018); Roy et al. (2019) observed, this aversion is reduced when the user has some level of control over the prediction process.

This situation illustrates that the successful integration of facial recognition systems in practical settings necessitates more than just technological progress. According to the AI Act by the European Commission (2024), the implementation of facial recognition should be proportionate and deployed only when strictly necessary for very specific cases. Negri et al. (2024) present a framework aimed at determining whether a facial recognition intervention is appropriate for a particular usage scenario. Other factors such as the application context, including the prospective end-users and the demographic characteristics of the population on which the system will operate, must be thoroughly taken into account. These approaches are closely connected to investigating human-centered ML techniques (Papenmeier et al., 2022), such as human oversight strategies for assessing and enhancing system outcomes (Hupont et al., 2022; Kyriakou and Otterbacher, 2023), as well as mechanisms for preserving the essential human element in decision-making in areas where safeguarding fundamental rights is particularly crucial, as highlighted by Koulu (2020).

In some studies, the preservation of the human element in decision-making is achieved by incorporating human factors into ML systems,

especially in tasks where distinctive human characteristics are found. Makino et al. (2022) found that in medical diagnosis on breast cancer screening, algorithms utilize features that experts often ignore and are outside areas that they considered suspicious. While it is desirable to have greater consideration of human factors in the automated decision-making process, researchers such as Hupont and Fernández (2019) have also studied the negative consequences of mimicking certain human biases. The *other-race effect* for face recognition (our ability to best recognize the identity of faces from our own race) has been observed in several human studies by Meissner and Brigham (2001); Feliciano (2016). Phillips et al. (2011) showed that the other-race effect may be present in algorithms.

There is a recent line of research investigating how human annotators can be effectively introduced into the loop so the algorithm can pass the final decision to the human when certain conditions are given (Hemmer et al., 2023; Mozannar et al., 2023; Keswani et al., 2021). These conditions are often related to the low confidence of an automated system, which can be used to determine what type of human-machine interaction is most appropriate in a hybrid system (Punzi et al., 2024), as well as to distinguish which annotations flows should be adopted to make human-machine collaboration more efficient (Lee et al., 2022). Combining decisions of systems and humans based on their perceived individual similarities has also been investigated by Phillips et al. (2018).

To the best of our knowledge, most of the efforts in integrating human factors into technology have mainly focused on encoding specific human traits and enhancing model performance – observing humans and refining models independently. Some more recent efforts have gone further, proposing novel techniques to combine human and machine performance. However, there is still a lack of understanding of the key differences of decision-makers, especially in contexts where the task involves a certain subjectivity. In a scenario where there is no

longer only the final decision related to the task at hand, but also the decision as to which agent – human, algorithmic or combination of both – should decide, it is important to know the strengths and weaknesses of both agents, which of these are shared and which diverge, and how these differences and similarities can be exploited. Here, we propose to study human and machine similarities and differences to capture and understand their complementary aspects so that this knowledge enables efficient and accountable human-machine interaction paradigms. For this purpose, we establish an error-centred comparison of human and machine performance in solving a face matching task. Examining these distinctions and similarities is crucial for enhancing the effectiveness of human oversight of algorithms, and for gaining insights into integrating the human factor into decision-making processes. Additionally, we explore how gender and ethnicity can influence human errors, building on earlier findings that have identified their significance in tasks that involve facial recognition (Phillips et al., 2011; Wright and Sladden, 2003).

4.3 Research Questions

The main goal of this work is to compare model errors with human errors in face recognition. We also investigate how human conceptions of gender and ethnicity affect these errors.

Our experimental setting, described in detail in §4.4, is based on a number of face recognition tasks that are performed by two automated systems, as well as by human annotators hired through a crowdsourcing platform. These tasks consist of matching facial images: given a pair of facial images, determining whether they belong to the same individual or to two distinct individuals. Both the two automated models and the set of annotators perform this task independently.

Error consistency We would like to characterize similarities and differences between human errors and ML system errors. To achieve this, first we need to determine if errors and successes are consistent, *i.e.*, if we can determine which are the subsets of face recognition tasks in which errors and successes are concentrated.

RQ1a *Are humans consistent in a face recognition task?*

RQ1b *Are ML systems consistent in a face recognition task?*

Error alignment We want to uncover whether there are common difficulties between ML systems and human annotators. We expect these common difficulties to manifest as incorrect human annotations on those face recognition tasks where the ML system erred. We also expect to obtain more incorrect annotations in cases where more than one ML system errs. If human annotators and ML systems provide a low-confidence annotation, then we would like to test whether their confidence in annotation aligns.

RQ2a *Are human annotators more likely to make a mistake in a face recognition task if a ML system also gives an incorrect answer, compared to tasks for which the system is correct?*

RQ2b *Are human annotators even more likely to make a mistake if more than one ML system is incorrect?*

RQ2c *Are human annotators' perception of similarity and ML computations of similarity correlated?*

The role of gender and ethnicity in errors False positives pairs involving people of different gender and/or ethnicity are unlikely to be made by humans, as differences related to gender expression and ethnic appearance are determining factors in humans when establishing an identity judgment (Phillips et al., 2011; Wright and Sladden, 2003). We would like to investigate whether cases where images are

perceived to have different gender/ethnicity by humans lead to lower confidence by a ML system. We remark that "human perception of gender and ethnicity," refers to differences and similarities in terms of gender expression and ethnic appearance, and not in terms of gender identity or self-ascription to an ethnicity.

RQ3a *Are ML errors on pairs labeled as depicting different gender expression, or eliciting different perceptions of ethnicity unlikely to be made by human annotators, to the extent that we can characterize "non-human-like" errors?*

RQ3b *Are human perception of similarity and/or ML similarity score correlated with human perception of gender and/or ethnicity similarity?*

Error-based human-machine collaboration We want to investigate if we can develop a strategy to optimise human-machine collaboration in the context of solving a face recognition task. The consistency raised in the first question would allow us to generalise in this context, while the study of the alignment between human and machine error patterns would allow us to detect the key points of complementarity for the development of a successful strategy.

RQ4 *Can we design a novel human-computer collaboration strategy based on the results of our comparative study?*

4.4 Experimental Setup

Our experimental setting is based on a number of face recognition tasks that are performed by two automated systems, as well as by human annotators. Given a pair of facial images, the task consists of determining whether they belong to the same individual or not.

Both the two automated models and the set of annotators performed this task independently.

4.4.1 Datasets

Training data We used two pre-trained face recognition models. Both were trained by their respective authors on *MS-Celeb-1M* (Guo et al., 2016), a dataset released by Microsoft in 2016. According to its authors, it was the largest publicly available face recognition dataset in the world. It contains about 10M images of nearly 100K people. After an investigation by Financial Times in 2019,[3] Microsoft removed MS-Celeb-1M. Before its demise, the dataset was widely used and still exists in several forms, such as trained models. MS-Celeb-1M is fairly unbalanced demographically (see Table 4.1).

Testing data We used *DemogPairs* (Hupont and Fernández, 2019) as the evaluation dataset. It contains 10,800 facial images corresponding to 600 people divided into 6 balanced demographic labeled folds: { female, male } × { Asian, Black, White } (see Table 4.1). These labels were manually annotated by their authors. We will use *labels* when we refer to those from the original dataset, and *annotations* for those obtained from our user study. DemogPairs was created and released by its authors with the explicit objective of being used as a tool to test for demographic biases on face recognition models.

4.4.2 Models

The two face recognition models used in this work were IR50+ArcFace (Deng et al., 2019) and LightCNN (Wu et al., 2018),

[3] *Who's using your face? The ugly truth about facial recognition* Financial Times, 18 September 2019.

	MS-Celeb-1M	DemogPairs
# Images	10M	10.8K
# People	100K	600
% Female	≈80%	50%
% Male	≈20%	50%
% White	76.3%	33.3%
% Black	14.5%	33.3%
% Asian	6.6%	33.3%

Table 4.1: Features of MS-Celeb-1M (Guo et al., 2016; Wang et al., 2019) and DemogPairs (Hupont and Fernández, 2019).

both trained over MS-Celeb-1M by the respective authors. We did not conduct additional training or fine-tuning for this work. Both pre-trained models can be found online (Wang et al., 2021; Wu et al., 2018).

IR50+ArcFace is an extension of ResNet50 (Guo et al., 2016; He et al., 2016), a residual network that has been extensively applied to many image tasks, with an ArcFace loss function (Deng et al., 2019). It reaches an accuracy of 99.78% in LFW (Huang et al., 2008), 97.53% in AgeDB (Moschoglou et al., 2017), 95.22% in VGGFace2 (Cao et al., 2018), well-known public benchmarks for pair matching. **LightCNN** was created to learn a compact embedding on large-scale face data with noisy labels. It has been reported to achieve state-of-the-art results on various face benchmarks without fine-tuning (Wu et al., 2018). In this work, we used the 29-layer version, which reaches an accuracy of 99.40% in LFW.

For evaluation, we used `face.evoLVe` (Wang et al., 2021), a face recognition library for face-related analytics and applications. For the purposes of this research, the library was instrumented to keep track of individual errors. The instrumented library is available with

our code release.

4.4.3 Methodology

We performed an online user study using oTree (Chen et al., 2016), with the following structure.

Participant recruitment We recruited participants through the crowdsourcing platform Prolific.[4] We considered four countries in continental Europe in which Prolific has large user bases: France, Germany, Italy, and Spain, plus the United Kingdom and Turkey. The crowdsourcing platform provides gender information and allows users to self-identify with a "simplified ethnic group" (Asian, Black, and White).We made sure that our sets of participants were gender balanced, and that for each pair of images, one person from each simplified ethnic group participated in their evaluation. So, for every pair of images, we collected 3 annotations. In total, we recruited 235 participants, excluding 2 of them from our data due to failed attention checks. For the subsequent analysis, based on ethnic self-identification, we selected 162 participants. Participants were paid 0.70 GBP (about 0.82€) to label 10 pairs of images, with an average completion time of 5 minutes. This amounts to 8.4 GBP per hour, which is slightly above the recommended payment by this platform (8 GBP/h).

Demographic questionnaire Participants were asked about their age, gender identity, and ethnic background (see Figure 4.1).

Face recognition tasks Participants evaluated one pair of images at a time. The participant had to answer the question *Are they the*

[4] www.prolific.co

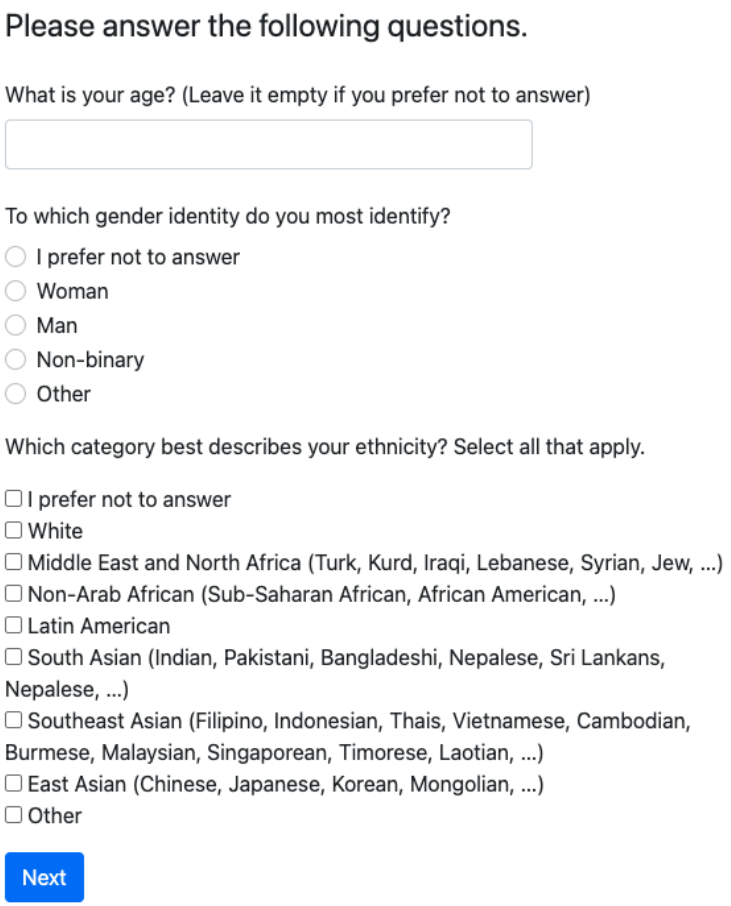


Figure 4.1: Participants were asked about their age, gender identity and ethnic background before starting with the face matching tasks.

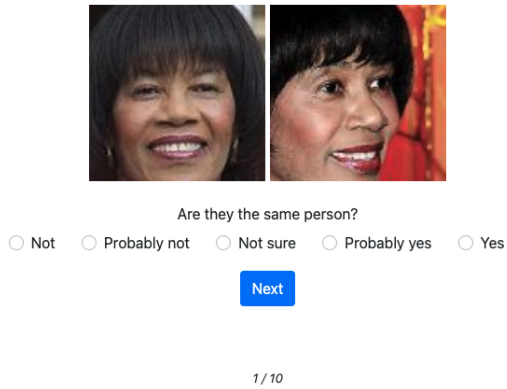


Figure 4.2: Pair in a face matching task.

same person?, with the possible options: *No*, *Probably not*, *Not sure*, *Probably yes* or *Yes* (see Figure 4.2). If the answer was different from *Yes*, then the same pair of images was shown one more time, and the participant was asked about some of the differences between the two images. These differences referred to gender expression, ethnic appearance, and age appearance (see Figure 4.3). The participant had to answer three questions: *How are these persons in terms of {gender expression | ethnic appearance | age appearance}*. Each question had to be answered independently on a scale with five options: *Different*, *Probably different*, *Not sure*, *Probably equal* and *Equal*. We remark that we asked about "expression" and "appearance" because the participants do not know the identities of the photo subjects.

Task selection We found that the joint accuracy of the face recognition models (see §4.4.4) was correct above 93% of the tasks. Hence, due to budget constraints, we annotated all the cases where the models were wrong ("misses"), and a sample of cases in which both models were right ("hits"). We annotated 363 "misses" (237 false negatives and 126 false positives), which were shown to a total of 164 participants, from which we selected a demographically balanced set of

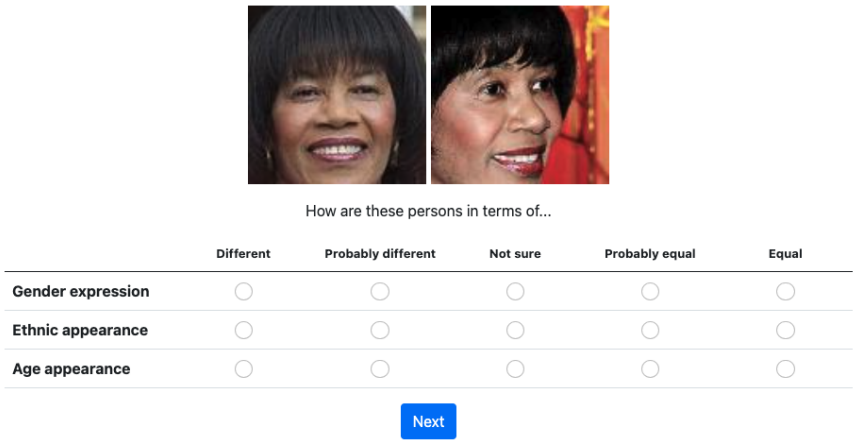


Figure 4.3: For those pairs where the participant was not completely sure of both identities being the same person (answering something different to *Yes* to the question *Are they the same person?*), participants were asked to provide some details relate to gender expression and ethnic appearance similarities.

108 participants. We also annotated 180 model "hits", which were shown to a total of 69 participants, from which we selected a demographically balanced set of 54 participants. This selection of "hits" was a random sample that was demographically balanced for the true positive set (90 pairs) and for the true negative set (90 pairs).

4.4.4 Measurements

We measured the following dependent variables:

Accuracy Accuracy is defined as the fraction of correct responses with respect to the ground truth.

1. **Machine accuracy**: Joint accuracy of the models. Given a pair, we calculate the average of the calibrated similarity scores of the two models, and the label is decided based on this average.

2. **Human accuracy**: Accuracy of the human annotators, as a group of three annotators. This is computed as a macro average, *i.e.*, first the three evaluations on a pair are averaged, and then we determine whether that average is correct or not.

Similarity This is a measurement of how similar the model or the human annotator perceives the persons in the images.

1. **ML similarity score**: Given two images, the model computes two embeddings. The normalized distance between these embeddings, d, is compared against a threshold θ to determine the output (if $d < \theta$, the pair is labeled as positive, negative otherwise). We take the calibration of $1 - d$ as the similarity of the pair, that can be interpreted as a probability. Scores close to 0.5 can be interpreted as a low model confidence.

2. **Human perception of similarity**: this is inferred from the annotator's actual answer to the questions in the face recognition tasks (see §4.4.3). From this measurement we can infer human confidence: the answers in the extremes (*No* and *Yes*) correspond to the highest confidence, while answer *Not sure* corresponds to the lowest confidence.

4.5 Results

In what follows, we will consider a *human error / success* when the mean response of the three annotators solving the same task corresponds to a incorrect / correct response, respectively. Similarly, we

will consider a *machine error / success* when the joint evaluation of the two models is incorrect / correct, respectively. For brevity, we will use "false positives", "false negatives", "true positives" and "true negatives" when we refer to the responses given by the models. In case we refer to the annotators' responses, we will do so explicitly. We will also show some results of significance test (p-values, noted as p). All these tests correspond to Kolmogorov-Smirnov tests.

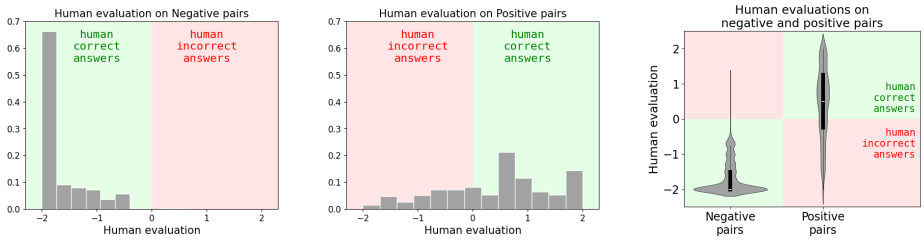
Participant Demographics Participants were on average 27.3 years old (SD=12.0 years). Out of the participants that indicated their gender, 55% identified as female, 41% as male, and 4% as non-binary. The majority of the participants that indicated an ethnicity identified as "White" (46%), followed by "Non-Arab African" (19%), "South Asian" (13%), and "East Asian" (9%). The remaining ethnicities accounted for less than 5% of the participants each.

Error Consistency (RQ1) We now consider the agreement of human annotations, *i.e.*, the extent to which multiple people agree on whether a pair of images represents the same person or not. Annotators were shown a total of 543 pairs of face images: 363 machine errors and 180 machine successes. Since the successes shown to the annotators are only a sample of all the successes from the models, we oversampled them to balance the workload. We also transformed every human annotation, originally based on a numeric 5-point scale, into a binary annotation in order to establish a fair comparison between human and machine agreement. We obtained a *moderate* multi-rater agreement among annotator responses (Fleiss' kappa = 0.47), which suggest that there is a mixture of agreement and disagreement between annotators (RQ1a). The moderate agreement present among the annotators is mainly due to the agreement they reach in those cases where the machine successes (Fleiss' kappa = 0.51), while in the cases where the machine makes a mistake we find no better agreement than would be the case by chance (Fleiss' kappa = -0.05). As

Figure 4.4c shows, human annotators are almost always correct in negative pairs, *i.e.*, when both images represent different people, as less than 5% of pairs are incorrectly classified as positive by the annotators. However, when images represent the same person, results are mixed. Although most of the positive pairs were correctly classified by the annotators, approximately 30% of those pairs were incorrectly categorized as negative. Differences in the distributions of labels on negative and positive pairs are significant at $p \ll 0.0001$.

For the models, we obtained an *almost perfect* inter-rater agreement between the outputs of IR50 and LightCNN (Fleiss' kappa = 0.92), which suggests that the agreement between models is much better than would be expected by chance (RQ1b). The human tendency to err with higher probability in positive pairs is similar to the models' way of erring: more than 65% of model errors are false negatives. The agreement among human annotators becomes significantly lower when we consider only *human errors* (Fleiss' kappa = -0.05), suggesting a poor agreement among annotators when their mean outcome is erroneous. This reduction in agreement is even more pronounced with the inter-rater agreement between models for *machine errors* (Fleiss' kappa = -0.29), which suggests a great disagreement in tasks where at least one of the models made a mistake. The interpretation of negative values for Fleiss' kappa are based on Landis and Koch (1977).

Error Alignment (RQ2) Next, we studied the extent to which human successes/errors are aligned with machine successes/errors. We considered four categories of model outcomes: True Positives, False Negatives, True Negatives, and False Positives. Human performance when evaluating True Negative pairs was significantly different from human performance when evaluating False Positive pairs ($p \ll 0.0001$). Differences were also significant in the case of human performance in the two subsets of positive pairs ($p \ll 0.0001$).

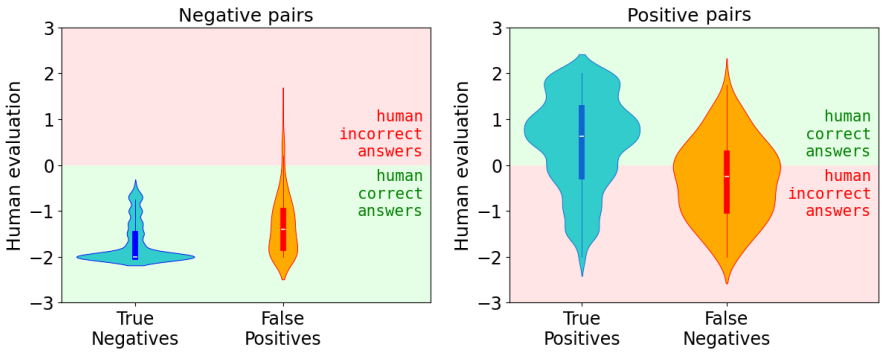


(a) (b) (c)

Figure 4.4: Human evaluations over 543 pairs of facial images. Negative pairs (a) correspond of images of different people (216 negative pairs), while positive pairs (b) correspond to images of the same person (327 negative pairs). On the right, we see the comparison of the distribution of negative pairs evaluations and positive pairs evaluations. Responses range from -2 ("No") to +2 ("Yes").

In the case of negative pairs, in which human annotators are almost always correct, Figure 4.5a shows less certainty and a possibility of error in the pairs in which ML models make a mistake. Human annotators are less likely to select the option "No" and more likely to select the option "Probably not" when asked about a pair of images of different people for which the ML models mistakenly indicated that they were the same person. In the case of positive pairs, shown in Figure 4.5b, we see a similar trend. In this situation, human errors are concentrated in the cases in which the models also made an error. In other words, there are some pairs of images of the same person for which both human annotators and models are likely to err. This, put together with the significance above, suggests that humans find cases where the machine erred more difficult in comparison to those where the machine succeed (RQ2a).

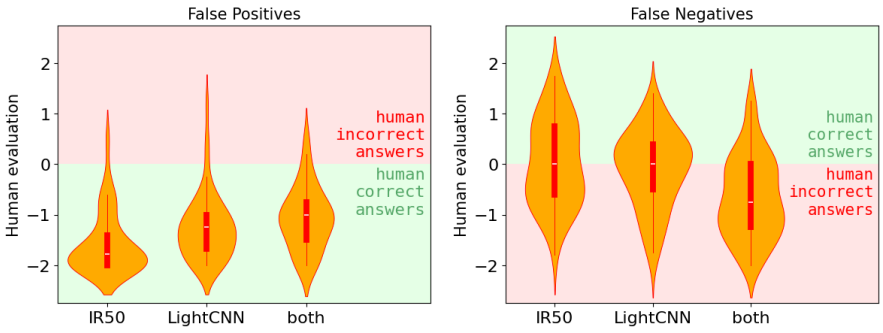
In general, annotators were more likely to make a mistake on pairs in which both models made an error (RQ2b); with human certainty



(a) (b)

Figure 4.5: Human evaluations over 543 pairs of facial images, from which 216 pairs are of different people (a) and 327 pairs of the same person (b). Human evaluations of machine errors (in orange, right violin in every figure) and successes (in blue, left violin in every figure). False Positives indicates pairs that the models mistakenly labeled as the same person. False Negatives indicates pairs that the models mistakenly labeled as different people. Human responses range from -2 to +2 ("No" to "Yes", respectively, to the question "Are they the same person?").

(preference for "No" over "Probably not") reduced in false positives of both models, and human error more likely in false negatives of both models. In the case of False Positives, human evaluation over those errors committed solely by IR50 are significantly different from human evaluations over those committed by both models, at $p < 0.001$. However, human evaluations over False Positives committed solely by LightCNN is not significantly different from human evaluation over False Positives committed by both models.We depict these differences in Figure 4.6a. In the case of False Negatives, human evaluation over those committed solely by IR50 is significantly different from human evaluation over those committed by both models ($p < 0.001$). Similarly, human evaluations over False Negatives committed solely



(a) (b)

Figure 4.6: Human evaluations of machine errors made by model IR50 only, errors made by model LightCNN only, and errors made by both models. Responses range from -2 ("No") to +2 ("Yes").

by LightCNN are significantly different from human evaluation over False Negatives committed by both models ($p \ll 0.0001$). We depict these differences in Figure 4.6b.

We examined human annotators' perception of similarity and compared them with model-computed similarity scores. This time we distinguished between eight overlapping categories of human and model errors and successes: { Human, Machine } × { True Positives, False Positives, True Negatives, False Negatives }. When both models and annotators gave correct responses, there were differences between the machine similarity score and annotator's perception of similarity (see blue violins in Figure 4.7). We found significant differences between both similarities for positive cases ($p \ll 0.0001$), and for negative cases ($p \ll 0.0001$).

This analysis reveals differences in the distribution of machine similarities, which tend to be bimodal and concentrated on the extremes, while human perceptions of similarity are more dispersed (RQ2c).

We found significant differences when both models and annotators

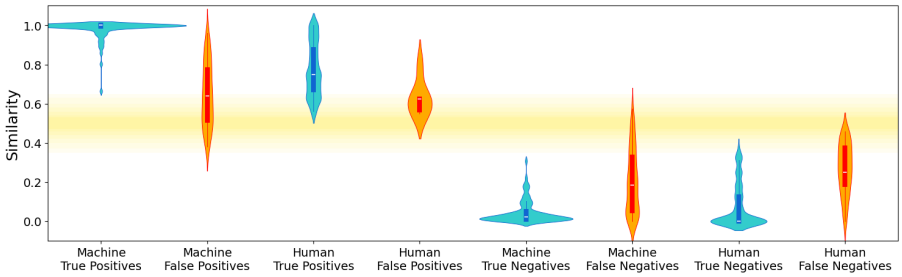


Figure 4.7: Human annotators perception of similarity and machine similarity score for different categories of human/machine errors (in orange) and successes (in blue). Note that machine confidence can be inferred from the similarity score (the further the similarity is from 0.5, the higher the confidence). The yellow band near a similarity score of 0.5 includes machine errors that can be anticipated as possible errors.

gave incorrect responses (see orange violins in Figure 4.7) for negative pairs ($p \ll 0.0001$), but not for positive pairs ($p = 0.35$). In the case of False Positives, annotators' perception of similarity when claiming a negative pair as positive tended to accumulate close to 0.5, indicating a low confidence in their answers (for comparison with machine similarity scores, human similarity 0.5 corresponds to the case "Not sure"). The yellow band around similarity 0.5 in Figure 4.7 includes machine errors that based on these observations could be predicted in advance as potential errors.

The Role of Gender and Ethnicity (RQ3) We examined human evaluations over two categories of False Positive errors, *i.e*, cases of the different people mistakenly identified by a model as being the same person. We compared pairs of facial images annotated as different in terms of gender expression **or** ethnic appearance, against pairs of facial images annotated as similar in gender expression **and** similar ethnic appearance.

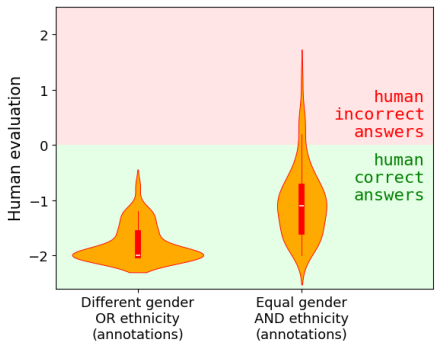


Figure 4.8: Human evaluation over False Positive machine errors. We examined pairs of images annotated as having different gender expression **or** different ethnicity appearance, compared to pairs annotated as having similar gender expression **and** similar ethnicity appearance. This indicates that annotators are less confident in differentiating between two distinct identities when they observe similarities in terms of gender expression and ethnic appearance.

Figure 4.8 shows the results, indicating that humans are less certain about their answer to the question on whether both images depict the same person (*i.e.*, more likely to indicate "Probably not" and less likely to indicate "No") for those pairs annotated as having equal or similar gender expression and similar or equal ethnicity appearance (RQ3a). Differences are significant at $p \ll 0.0001$.

People state that images of the same person portray the same gender expression and ethnic appearance. This is more evident (less noisy) in the case of gender expression, suggesting that this signal determines more directly the human label than ethnic appearance. Another possible explanation is that ethnic appearance might be more affected by different lightning conditions in the images (RQ3b).

In our experiments, we found partial evidence of the "other-race" effect (see Table 4.2). We calculated the error rates for the three self-ascribed ethnicities: White, Black, and Asian. We considered only

Figure 4.9: Distribution of human perception of gender expression similarity (1 - Different, 5 - Equal) for different human and machine outcomes, compared to human similarity perception, and machine similarity score, respectively.

pairs of images with the same ethnicity label in both images and computed the error rate for each of these sets of pairs. "White" and "Black" annotators are the most accurate when annotating images of their same kind, but this was not the case for "Asians". No "other-race" effect was found in the models.

4.5.1 Exploratory study of error-based human-machine collaboration

To study our fourth research question, we conducted a study with the intention of illustrating the consequences of applying a human supervision strategy based on the results previously obtained. We studied the improvement over model accuracy that would result from manually reviewing the pairs evaluated by the machine. The joint accuracy achieved by both models was 93.5%.

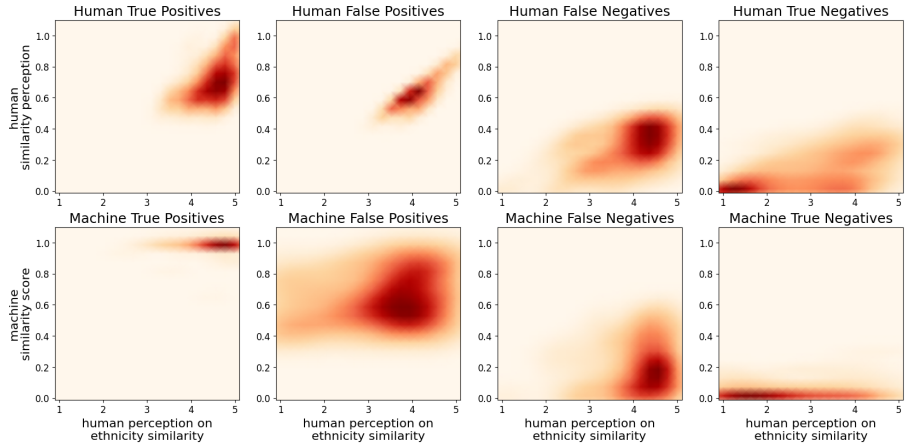


Figure 4.10: Distribution of human perception of ethnicity appearance similarity (1 - Different, 5 - Equal) for different human and machine outcomes, compared to human similarity perception and machine similarity score, respectively.

We apply this supervision strategy in a "human as overseer" scenario, in which each prediction made by the model is, in turn, made by the human, who confirms or corrects it in favor of their own prediction. Since the processing speed of the model is faster than human supervision, a queue of model predictions waiting to be supervised would be generated (see Figure 4.11). In this queue, we apply two rearrangements based on the insights previously explained.

The first rearrangement is based on the results obtained related to RQ2c: the use of machine confidence to prioritize those cases that have a high probability of being corrected by the human annotator. Note that machine confidence can be inferred from the similarity score (the further the similarity is from 50%, the higher the confidence, see Figure 4.7). The evolution of joint accuracy when this prioritization is implemented can be seen in the top pink line of Figure 4.12a. We can observe that the pairs that human annotators are able to solve correctly are concentrated at the beginning of the workflow, leading

	White-White pairs	Black-Black pairs	Asian-Asian pairs
White	0.10	0.45	0.20
Black	0.55	0.04	0.02
Asian	0.18	0.06	0.09
Machine	0.09	0.07	0.09

Table 4.2: Human and Machine error rate. First three rows are demographic groups evaluating different set of pairs. "White-white" pairs stands for pairs containing images of two people labeled as white, and so forth.

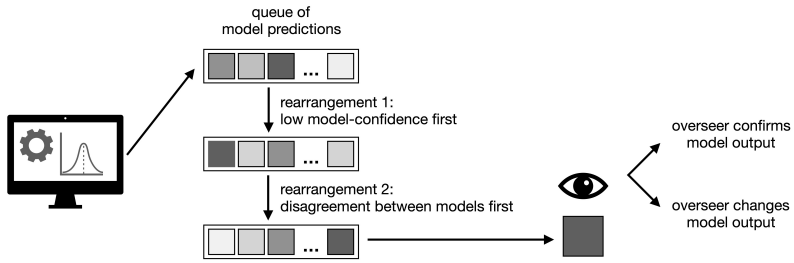
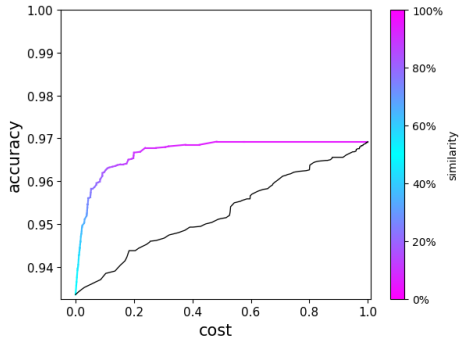
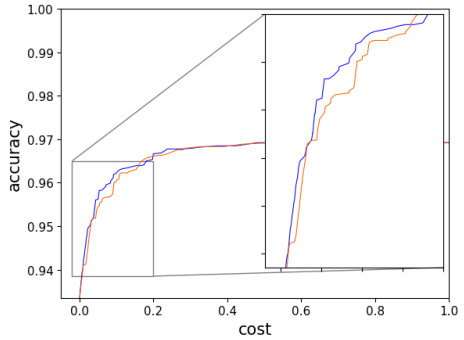


Figure 4.11: "Human as overseer" supervision strategy involving reorganization of the model's prediction queue.

to an early and rapid growth of the joint accuracy. This marked improvement in accuracy contrasts with the results we would obtain if this strategy were not taken into account (see the black line in Figure 4.12a). The second rearrangement is based on the results obtained when investigating RQ2b: prioritizing those pairs where the models gave different answers, *i.e.*, only one of the two models correctly classified the pair. As we can see in Figure 4.12b, the joint accuracy obtained if these pairs are prioritized (blue line) exceeds the joint accuracy obtained when this priority is not implemented (orange line) during most of the human annotation flow.



(a)



(b)

Figure 4.12: Human evaluation of pairs classified by the machine. (a) shows the evolution of joint human-machine accuracy when annotations are done in increasing order of machine certainty. (b) shows the evolution of joint human-machine accuracy if, in addition, we prioritized those pairs where the models gave different answers (blue line). This accuracy exceeds the joint accuracy obtained when this priority is not taken into account (orange line). The initial machine accuracy was 93.5%. Cost represents the rate of the number of annotators.

4.6 Discussion

This study comparing human and machine errors allows us to develop strategies for enhancing the efficiency and effectiveness of human-computer collaboration in the domain of face recognition tasks. A fundamental key to these strategies is to take full advantage of human capabilities to complement machine deficits and vice-versa. Three main observations can be deduced from investigation of the first three research questions.

First, sufficient consistency was observed in both the annotations provided by the humans and the responses generated by the model. The consistency in human responses enables us to identify errors that models make but are unlikely to be made by humans. In our setting, these are false positive pairs. When examining the correlation between human and machine errors, we observed that when the machine errs, humans are more likely to make a mistake as well, particularly in the case of positive pairs, where humans have a high error rate in comparison with the machine false negative error rate. However, in negative pairs, although human confidence decreases for those cases where the machine fails, human responses are mostly accurate, suggesting that the machine struggles with certain negative pairs that humans do not find difficult. Humans have a significantly better capacity to distinguish negative pairs than machines. Of the 126 false positives made by the machine, humans made only 6 errors (4.8%).

Next, we categorize machine errors into those occurring in just one of the two models and those occurring in both models.We find that in the case of a machine false negative, humans are notably more prone to error when that error is committed by both models. This reveals a correlation between the challenges faced by both models and those faced by humans when evaluating positive pairs. This indicates that humans have a significantly better capacity to correctly classify those

pairs in which both models disagree, over those pairs in which they are wrong. However, when a human assesses a machine false positive pair, the probability of error is not significantly influenced by whether the error is common to both models or not, which is consistent with the observation above.

Finally, we observe that only when humans and machine make an error by failing to detect a positive pair, their similarity scores could become similar. In all other scenarios of correct and incorrect identifications, the human and the machine provide responses based on notably different similarity scores. Moreover, there is a substantial disparity between the machine's similarity rating for correct identifications and the machine's similarity score for incorrect ones. This difference is even more pronounced when the machine classifies a pair as positive. This, combined with the high accuracy mentioned above of humans over machines in evaluating negative pairs, could help anticipate potential errors and suggest that a manual examination of these cases by a group of annotators could be beneficial.

Based on these observations, following the suggested strategy, we could improve the accuracy of the system by approximately 3 percentage points by assuming 10% of the total cost only. This improvement of 3% is equivalent in our case to the correction of 148 pairs (98 negative and 50 positive) misclassified by the machine, which outweighs the improvement of just 0.4% (equivalent to 27 corrections, 19 negative and 8 positive pairs) that we would achieve if we did not follow the proposed strategies. This study is a clear example of how the findings of such comparative and exploratory analysis can facilitate the proposal of simple but powerful human-machine interaction paradigms.

4.7 Conclusions

The main conclusions drawn from this work are (1) the facial recognition models shows a marked disparity in similarity scores between correctly and incorrectly resolved pairs, (2) there is a correlation between the shared challenges faced by the models and the difficulties experienced by humans, whereas humans encounter fewer issues when classifying pairs where the models provided different results, and (3) humans perform substantially better than facial recognition models in assessing negative pairs.

Observation (1) enabled us to detect potential errors in the facial recognition models, while observations (2) and (3) helped us prioritize those potential errors that a human annotator has a high chance of correcting. Implementing this in practice allowed us to design a manual evaluation strategy that achieves maximum joint human-machine precision with a very low number of annotations.

It is worth paying attention to the impact that some of these conclusions could have on facial recognition tasks in real-world contexts. As we saw in the human-machine collaboration paradigm proposed above, most of the machine errors corrected by a human annotator are negative pairs that were predicted as positive by the model. This should be taken into account when evaluating the suitability of integrating an automatic face recognition system in the specific application domain. In scenarios where the occurrence of false positives might have serious consequences and potentially affect fundamental rights, if there is a lack of adequate human oversight, the integration of facial recognition technologies demands a rigorous and thoughtful reconsideration.

In use cases where the resolution of face recognition tasks by a machine learning system can be conveniently monitored by human reviewers, it is still imperative to implement oversight strategies that acknowledge and address the disparate error patterns exhibited by

humans and machines.

A noteworthy aspect is the observation that the human advantage over machines in assessing negative pairs might be linked to the perception that humans have built on notions of similarity and difference in gender expression and ethnic appearance. When the human correctly classifies a negative pair that was classified as positive by the machine, both perceptions of gender and ethnicity seem to play a distinctive role in the final human decision. This apparent human tendency to use gender and ethnicity-related characteristics to differentiate negative pairs could be due not only to gender and racial stereotypes perpetuated in society, but also to the predominant presence of stereotypical images in face recognition databases (Dominguez-Catena et al., 2022; Keyes, 2018).

Finally, our results suggest that facial recognition algorithms are not advanced enough to fully replace human roles in real world scenarios. This may also not be desirable, especially in light of the recent ethical and legal concerns that have been raised about the use of this technology. The AI Act (EU Regulation 2024/1689) by the European Commission (2024) contains many explicit and implicit allusions to facial processing, whose applications are considered at different risk levels, including *high risk* and *forbidden*. This envisions a future scenario for face recognition technologies in which permanent human oversight will be essential, highlighting the value of preserving human input in decision-making.

4.8 Ethics Statement

Regarding the gender and ethnicity discussed in this chapter, it is important to note two things: (1) the original labels regarding gender and ethnicity in the testing database were inferred by means other than directly asking the person in the image about their demograph-

ics, so they should in no way be assumed to be true, and (2) in our study the participants were shown a pair of images and were asked about the similarity of *gender expression* and *ethnic appearance*, these being different concepts to those relating to the social identities of people in the images.

5 The Influence of Task Difficulty and Machine Accuracy

This chapter is based on Estévez-Almenzar, M., Baeza-Yates, R., and Castillo, C. *Human Response to Decision Support in Face Matching: The Influence of Task Difficulty and Machine Accuracy* (HHAI 2025); and Estévez-Almenzar, M., Baeza-Yates, R., and Castillo, C. *Brief Onboarding Phase Improves Decision Support: Evidence from Two High-Risk Scenarios* (Submitted for publication)

5.1 Introduction

Decision support systems (DSS) are a key modality of use of Artificial Intelligence (AI). They can be categorized by the extent to which the final decisions depend on them, from having no influence whatsoever, to being fully autonomous, with most cases operating somewhere between these two extremes (McGee et al., 1998; Cummings, 2017). Ideally, these systems together with human operator(s) create a hybrid human-machine intelligence that exploits the expertise of human operators with the capacity to find patterns in historical data that yield better decisions. This is particularly critical in applications that have significant effects on people, such as those described as *high risk* by the European AI Act (EU Regulation 2024/1689 –

European Commission (2024)). In these applications, given the ethical and legal requirements for human oversight, it is unlikely that fully automated systems are deployed in the near future. Instead, hybrid systems combining human and machine intelligence are likely to become the norm. When humans and machines work together, they should be evaluated together (Matias, 2023). However, human-algorithmic behavior involves complex emerging patterns, uncertainties, and a certain degree of unpredictability that is in tension with the goal of developing safe and trustworthy systems. Modeling how human operators respond to recommendations produced by an algorithm is paramount. This is an active research topic, and previous work investigates aspects such as the nature of the task for which decision support is provided (Bonnefon et al., 2024; Mahmud et al., 2022; Castelo et al., 2019), the way in which machine assistance is framed (Hou and Jung, 2021), preconceptions that make users averse (Dong and Bocian, 2024) or overreliance (Bashkirova and Krpan, 2024) on algorithms, and the accuracy or perceived accuracy of the algorithmic suggestions (Hou and Jung, 2021), among other factors.

In this chapter, we incorporate an additional use case to the one studied in the previous chapter. In this chapter we also incorporate human-machine interaction that was absent in the previous chapter. So, we study two scenarios: 1. face matching, in which a person is tasked with determining whether two photos correspond to the same person, and 2. criminal recidivism prediction, in which a person is tasked with determining whether a prison inmate for which parole is being considered will commit a new violent crime if released on parole. In both scenarios, people are assisted with a DSS. Our work addresses a series of research questions related to the extent to which a DSS can enhance human accuracy, focusing on responses to both correct and incorrect suggestions and the effects of fluctuating (increasing or decreasing) system accuracy. The addressed questions are the following.

RQ1 *Does an AI-based decision support system improve human per-*

formance?
We test to what extent the support of a high-accuracy machine improves human accuracy, while testing whether this improvement depends on the difficulty of the task assigned.

RQ2 *Does a low accuracy AI-based decision support system improve or deteriorate human performance? Does a misleading AI-based decision support system improve or deteriorate human performance?*
We test to what extent the support of a low-accuracy machine and a misleading machine deteriorates human accuracy, while testing whether this deterioration depends on the difficulty of the task. By *misleading machine* we mean a machine that is consistently producing a bad recommendation – we envision a machine like this can be in operation either due to an error during deployment, or a malicious action.

RQ3 *Does a variable accuracy AI-based decision support system improve or deteriorate human performance?*
We test to what extent the support of a variable-accuracy machine improves or deteriorates human accuracy, while testing whether this change depends on the difficulty of the task. We also test whether this change depends on human awareness of this machine variability. By a variable-accuracy machine we mean a machine that undergoes one or more changes during the interaction flow, affecting the accuracy of the suggestions – we envision a machine like this can be in operation either due to system updates, a change in the data distribution, or a malicious action.

We tackle these questions through a series of experiments (§5.3) conducted via crowdsourcing in which experimental variables include task difficulty, decision support accuracy, and whether a notification is given to users when decision support accuracy might change. Our results (§5.4) and discussion (§5.5), show a strong influence of task

difficulty, which not only makes human annotators less accurate, but also makes them more prone to be misled by inaccurate decision support and less aware of the accuracy of different decision support systems. We also find important differences on human perceptions when decision support errors appear at the beginning or end of the sequence of tasks, compared to when they are randomly distributed throughout the sequence of tasks. The last section (§5.6) presents our conclusions and outlines future work.

5.2 Decision Support in High-Risk Scenarios

In addition to the reviewed literature presented in Chapter 2, here we review further research related to this chapter.

Human-DSS Decisions in Face Recognition Human-AI collaboration in face recognition tasks presents both opportunities and challenges, particularly due to the significant variability in human face recognition abilities. Young and Burton (2018) show that the ability to recognize unfamiliar faces varies widely across individuals and is only modestly influenced by training or occupational exposure. This variability, studied by Ramon (2021), has led to the identification of so-called "Super-Recognizers" (SRs): individuals with exceptional performance in unfamiliar face matching and memory tasks. Such individual differences are not only theoretically informative but also practically significant, as White and Burton (2022) suggest that the effectiveness of human-AI systems may depend heavily on the specific human users involved.

Recent work by Towler et al. (2023) highlights the distinct cognitive and perceptual profiles of SRs, trained forensic examiners, and

deep neural networks, even when all achieve similar levels of accuracy. These differences manifest in error patterns, decision strategies, and even underlying face representations, emphasizing that human and machine recognition are not interchangeable but complementary. However, collaboration between humans and AI in face recognition is not without risk. Fysh and Bindemann (2018) observed that prior decisions made by face recognition systems influenced subsequent face matching decisions made by human operators. When face pairs were incorrectly labeled by the machine, the precision of humans decreased by drawing attention away from face images, even when humans were warned that machine predictions could be inaccurate. Furthermore, Salehi et al. (2021) observed that decision-deferral rates in human-machine systems influence both human performance and trust during face-matching tasks.

Jeckeln et al. (2018) documented that human face recognition accuracy can be improved by the wisdom of crowds: combined judgment of many is better than the decision of an individual. Ranjan et al. (2019) observed that there is a similar benefit to merging the performance of multiple algorithms. When considering decisions resulting from the fusion of human decision and machine decision, the results in Phillips et al. (2018) lead to large performance improvements compared to the human response or the algorithm response alone.

Human-DSS Decisions in Criminal Recidivism Prediction Human-AI collaboration in criminal recidivism prediction has emerged as a focal point in debates over fairness, accuracy, and the proper role of algorithmic tools in high-stakes decision-making. Risk assessment instruments (RAIs) are now widely used by judges, parole boards, and correctional authorities to inform decisions about incarceration, release, and supervision, based on the premise that these tools can outperform unaided human judgment (Monahan and Skeem, 2016; Neufeld, 2017). However, growing evidence complicates this assumption. The study by Dressel and Farid (2018) have shown

that laypeople with no criminal justice training can match or even outperform some widely used RAIs, such as COMPAS, under certain conditions. However, Lin et al. (2020) show that this parity vanishes when participants are deprived of feedback or when the system has access to a richer set of input features, scenarios in which algorithms tend to outperform humans. Similarly, Portela et al. (2024) observe that RAIs result in more accurate predictions by both expert and non-expert users, and highlight that expert participants, although they would not use a fully automated system in criminal risk assessment, do find it valuable for training.

Beyond accuracy, concerns about the fairness and social impact of RAIs are well-documented (Green and Chen, 2019; Angwin et al., 2022; Chouldechova, 2017; Kleinberg et al., 2016). For instance, Green and Chen (2019) revealed that the use of these tools can introduce or amplify racial disparities, particularly through "disparate interactions" where AI predictions appear to influence human decisions differently based on a defendant's race. Furthermore, participants often struggle to assess both their own and the algorithm's accuracy, and their reliance on AI may vary in problematic ways depending on perceived fairness or confidence. Green and Chen (2019) also advocate for reframing AI-assisted decision-making within a broader sociotechnical context – one that emphasizes improving human judgment rather than optimizing algorithmic performance in isolation. Others such as Chiang et al. (2023) have explored collective decision-making as a mitigating factor: group deliberation appears to support fairer outcomes and more confident rejection of flawed algorithmic advice, especially when groups correctly identify AI errors. These findings suggest that integrating AI into criminal justice decisions requires not only technological robustness but also careful attention to the human, cognitive, and institutional frameworks in which such tools operate.

This chapter addresses task complexity, a subject that has not been

explored in as much detail as other aspects in the surveyed literature, as Salimzadeh et al. (2023) stated, positioning it as a pivotal element to be carefully considered in the design of decision support systems for two high-risk scenarios: face recognition and criminal recidivism prediction. For the purposes of this study, we define complexity as the difficulty perceived by the human agent, thus prioritizing the human factor in a field that may involve high risks. Additionally, in many real-world scenarios in which data evolves, and given that machine learning models should be trained and applied on data with identical distributions, keeping models up to date is a critical task (Majidi et al., 2024; Faubel et al., 2023; Bayram and Ahmed, 2024). Updating the models produces variability in performance, causing an effect on interaction patterns (Renier et al., 2021). As far as we know, previous research has not thoroughly examined the effects of potential machine variability on human-machine interactions, nor has it been thoroughly studied the effects of how errors are distributed along a sequence of tasks in machines of equal average accuracy. The interplay between task complexity and machine variability has also received little attention. In this chapter we consider machines of varying accuracy and observe how these variations affect human performance. We also study whether notifying the human operator each time a variation in the machine occurs makes any difference in joint human-machine performance.

5.3 Study Design

To cover the research questions outlined above, we considered three independent variables: (i) task difficulty, (ii) machine accuracy, and (iii) change notifications. *Task difficulty*, described in §5.3.1, indicates the difficulty of the tasks assigned to the participant. *Machine accuracy*, described in §5.3.2, indicates the accuracy of the decision support assigned to the participant. *Change notification* indicates

whether the participant is notified or not when the decision support system changes, and only applies to those experiments in which the accuracy varies. We measured three dependent variables: accuracy, influence factor, and confirmation factor. The three dependent variables are defined in §5.3.3. We designed three experiments, each of them carried out in two high-risk scenarios: face matching and criminal recidivism prediction.

Experiment 1: With / Without Decision Support This experiment works as our "control experiment" and is designed for the purpose of investigating RQ1. Two groups of different participants were compared. The control group solved the task without machine suggestions. The experimental group of participants received, for every task, a suggestion from a system having a high (95%) accuracy.

Experiment 2: With Degraded Decision Support This experiment is designed for the purpose of investigating RQ2. Two experimental groups were compared. One of them solved the face matching tasks while receiving suggestions from a misleading machine (5% accuracy) and the other one while receiving suggestions from a low (50%) accuracy machine.

Experiment 3: With Variable Decision Support Machine This experiment is designed for the purpose of investigating RQ3. Six groups of different participants were compared. Two groups solved the tasks while receiving suggestions from an *increasing* accuracy machine (that we note INC machine), two other groups solved the tasks while receiving suggestions from a *decreasing* accuracy machine (DEC), and two other groups solved the tasks while receiving suggestions from a *temporary fail* machine (FAIL). For each machine, one of the groups was notified every time the machine changed, and the other group was not.

Face Matching										
	Experiment 1		Experiment 2		Experiment 3					
Machine	none	*high*	*misleading*	*low*	INC		DEC		FAIL	
Change notification	-	-	-	-	yes	no	yes	no	yes	no
# participants	20	20	20	20	20	20	20	20	20	20
Criminal Recidivism										
	Experiment 1		Experiment 2		Experiment 3					
Machine	none	*high*	*misleading*	*low*	INC		DEC		FAIL	
Change notification	-	-	-	-	yes	no	yes	no	yes	no
# participants	20	20	20	20	20	20	20	20	20	20

Table 5.1: Summary of experiments. Each experiment was carried out twice (once for each problem difficulty, "easy" or "hard").

5.3.1 Procedure

We performed an online user study, with the following structure.

Tasks selection For each application (face matching and criminal recidivism), each of the experiments was carried out twice. Once with a set of tasks that we call *Easy Set*, and once with a different set of tasks that we call *Hard Set*. Next, we explain how these sets were created for each application.

In the case of **face matching** tasks, both sets consist of pairs from the DemogPairs (Hupont and Fernández, 2019) testing database, and their classification as *easy* or *hard* is based on the study presented in Chapter 4, in which a total of 540 pairs were tested by a total of 162 participants (10 different pairs per participant, 3 different participants per pair). As explained more in detail in the previous chapter, in this study participants rated these pairs by answering the question "Are they the same person?" with the options "No", "Probably not", "Not sure", "Probably yes", or "Yes". This allows us to categorize some of these pairs according to the difficulty experienced by partici-

pants in solving the task. So, the so called *Hard Set* of face matching tasks is composed of tasks where participant responses present a low certainty and a high error rate. More specifically, we selected those pairs whose mean participant response was very close to "Not sure", and at least one of the three participants made a mistake. We also selected those pairs where the average response is exactly "Not sure". We obtained a total of 69 pairs (63 positive and 6 negative). After a careful manual inspection, from among the 63 positive pairs we chose the 24 most difficult ones which together with the 6 negative ones form the *Hard Set*. On the other hand, the so called *Easy Set* of face matching tasks is composed of tasks where participant responses had a low certainty and a low error rate. The reason for considering pairs with *low* participant certainty rather than *high* participant certainty is to avoid pairs that are trivial to the participant (mostly negative pairs, some with notably diverse gender expression or ethnic appearance, as seen in Chapter 4). So, the pairs in this set have a slightly more relaxed human certainty condition: we selected those pairs whose mean response was close to "Not sure", avoiding those whose mean was exactly "Not sure". From these pairs, we randomly chose 30 pairs of images, maintaining the above ratio of 24 positive and 6 negative pairs, and avoiding repetitions with the "Hard Set". This left us with 30 pairs that form the *Easy Set*.

In the case of **criminal recidivism prediction** tasks, both sets consist of a selection from a total of 90 semi-synthetic cases generated by Portela et al. (2024). These cases are based on an original dataset with 597 anonymized cases of people that have served a prison sentence were evaluated. To make the selection of easy and hard cases, we carried out a preliminary study, similar to the one for face matching tasks, where these 90 semi-synthetic cases were annotated by a total of 45 participants (10 different cases per participant, 5 different participants per case). Participants had to answer the question *What is the probability of this person to be arrested for committing a new violent crime in the next 2 years?*, with the possible options: *Level 1 (<5%)*, *Level 2 (5% - 29%)*, *Level 3 (30% - 49%)*, *Level 4 (50%*

- *85%)*, or *Level 5 (>85%)*. These are the risk levels used in their original paper Portela et al. (2024). The interface for the evaluation of each task in this preliminary study was the same as the one for the evaluation of the tasks in our experiments (see Figure 5.2). This preliminary study allows us to categorize some of these cases according to the difficulty experienced by the participants when solving the task. In this case, the proportion of recidivism cases is also taken into account when defining the *Easy Set* and the *Hard Set*. Our intention was to make it as similar as possible to the actual proportion of violent recidivism cases in Catalonia, which is the region of the source data. An executive report on prison recidivism rates in 2020 carried out by the Center for Legal Studies of the Generalitat de Catalunya (Capdevila et al., 2023) indicates that 21.1% of people who are released after serving their first sentence reoffend, but only 5% reoffend violently. So, we planned to include 2 violent recidivism cases and 13 non-recidivism cases in each set. So, the so called *Hard Set* of criminal recidivism prediction tasks is composed of tasks where participant responses present a low certainty and a high error rate. More specifically, we selected those cases whose mean participant response was very close to *Level 3 (30% - 49%)* and at least three out of five participants made a mistake. We obtained a total of 16 cases, all of them being non-recidivism cases. We randomly selected 13 of them and added 2 recidivism cases with a high error rate. The so called *Easy Set* of criminal recidivism prediction tasks is composed of tasks where participant responses present a high certainty and a low error rate. In this case, we did not worry about avoiding trivial cases, as we did in the facial matching tasks, since this degree of triviality does not exist when predicting criminal recidivism. So, the cases in this set correspond to those cases whose mean response was far from *Level 3 (30% - 49%)*, where nor more than two out of five participants made a mistake. We obtained a total of 19 cases, of which 6 were cases of recidivism. We randomly 2 of these 6 recidivism cases that together with the remaining 13 non-recidivism cases formed our *Easy Set*.

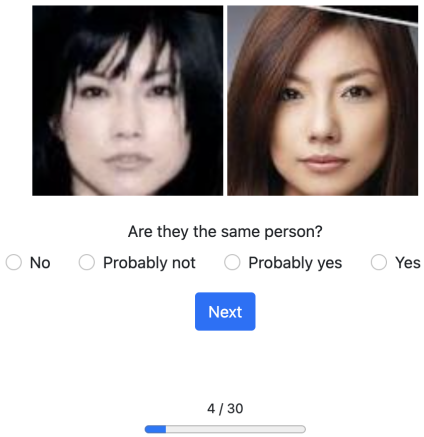


Figure 5.1: Interface for the evaluation of face matching tasks.

Participant recruitment As for the preliminary studies that allowed us to distinguish two sets of tasks with different level of difficulty, we recruited participants through Prolific.We considered four countries in continental Europe in which Prolific has large user bases: France, Germany, Italy, and Spain, plus the United Kingdom. We made sure that our sets of participants were gender balanced, a feature that the platform provides based on participants' disclosures of gender. For **face matching** tasks, we recruited a total of 400 participants, and each participant annotated 30 different pairs of images. In total, we collected 12,000 participant's annotations, and a total of 60 different pairs of images (30 *easy*, 30 *hard*) were annotated under various conditions. For **criminal recidivism prediction** tasks, we recruited a total of 400 participants, and each participant annotated 15 different cases of prisoners. In total, we collected 6,000 participant's annotations, and a total of 30 different cases of prisoners (15 *easy*, 15 *hard*) were annotated under various conditions. Participants were paid an average of 8.68 EUR per hour. Specifically, they were

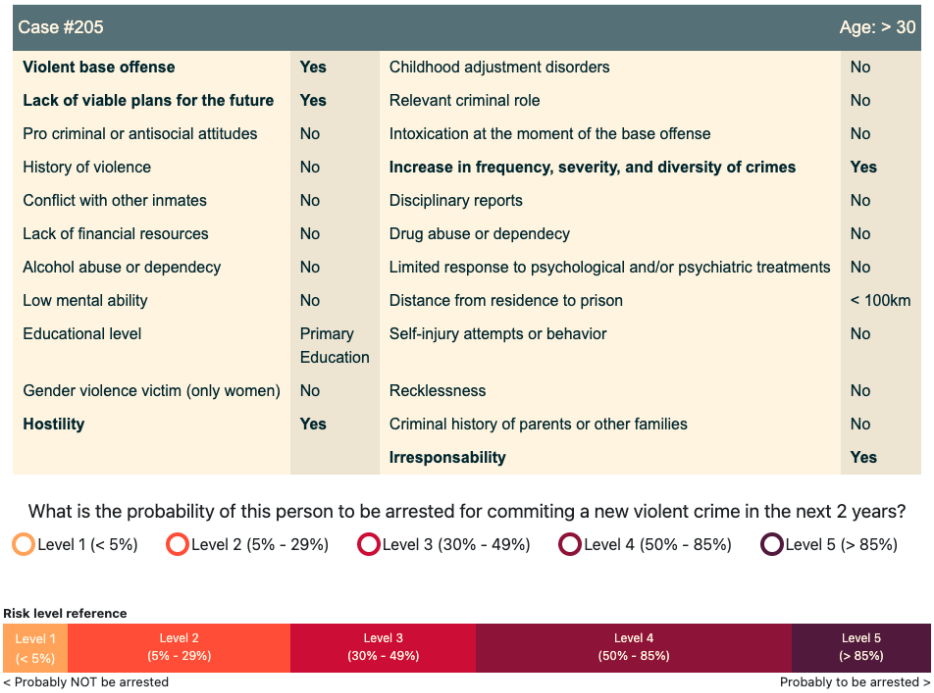


Figure 5.2: Interface for the evaluation of criminal recidivism tasks, representing the case of an imprisoned person. It consists of a list of 23 items that are used by RisCanvi to predict violent recidivism (Portela et al., 2024).

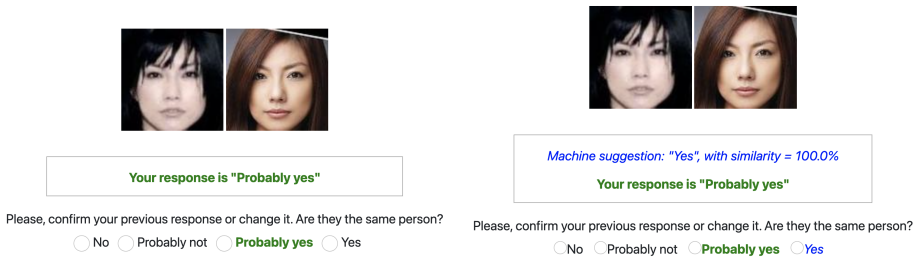
paid 1,07 EUR for labeling either 30 pairs of images in the case of face matching tasks, or 15 cases in the case of criminal recidivism, with an average completion time of 10 minutes. To encourage participants' effort, we set a bonus: participants were warned (and reminded during the study) that if they managed to correctly match more than 80% of the tasks they would receive an extra payment of 30%.

Evaluation of tasks Participants evaluated one task at a time. Attention-check questions were included to filter out inattentive par-

ticipants. In **face matching** tasks, given a pair p, the participant had to answer the question *Are they the same person?*, with the possible options: *No*, *Probably not*, *Probably yes*, or *Yes*. Here, the option *Not sure* was omitted to encourage participants to make a decision in one direction or the other. After answering, if the participant was assigned to one of the experimental groups, the machine suggestion was shown together with the machine similarity score associated with the pair s_p on which the suggestion was based: $0 \leq s_p \leq 0.25$ with suggestion *No*, $0.25 < s_p \leq 0.50$ with suggestion *Probably not*, $0.50 < s_p \leq 0.75$ with suggestion *Probably yes* or $0.75 < s_p \leq 1.00$ with suggestion *Yes*. They had the possibility to modify their answer (see Figure 5.3b). Participants in the control experiment also had the possibility to modify their answer (see Figure 5.3a). In **criminal recidivism prediction** tasks, participants were shown one semi-synthetic profile representing a person who served a prison sentence at a time (see Figure 5.2), consisting of a list of 23 items that are used by RisCanvi to predict violent recidivism (Portela et al., 2024). Participants had to answer the question *What is the probability of this person to be arrested for committing a new violent crime in the next 2 years?*, with the same possible options as in the preliminary study, described previously. After answering, if the participant was assigned to one of the experimental groups, the machine suggestion was shown together with the recidivism probability estimated by the machine (see Figure 5.4). All participants had the possibility to modify their answer, both those with a machine assigned and those without.

Exit survey After evaluating the tasks assigned, the participants who interacted with a machine completed an exit survey.

They were asked whether the AI: 1. *gave the participant good suggestions*, 2. *helped the participant find the right answer*, 3. *influenced the participant's final answers*, 4. *made the participant more confident*, 5. *made the participant more confused*, 6. *whether the participant felt confident about their own answers*, and 7. *whether the participant*



(a) Control group (b) Experimental groups

Figure 5.3: Screenshots from the evaluation of face matching tasks.

was satisfied with the machine (see Figure 5.5). This exit survey is based on previous work by Pu et al. (2011); Dietvorst et al. (2015). For each question, a five-point Likert scale was used (*Strongly disagree*, *Disagree*, *Neither agree nor disagree*, *Agree*, *Strongly agree*). This survey is proposed to study the perception of the participant about how useful the machine is perceived. So, for the analysis we discard question 5, that works as an attention check question when compared to question 4, and we also discard question 6, which is not related to how the machine suggestions are perceived.

5.3.2 Decision Support Accuracy

To simulate machines that adapt to the circumstances that we want to reproduce in each experiment, we add noise to a highly accurate model. This method is widely used in the literature on human-machine studies (Carragher and Hancock, 2023) and consist of applying noise to the probability p given by the model, where $0 \leq p \leq 1$. We define the noise as $f(p) = 1-p$, which gives us a noisy probability that forces the opposite response. The number of pairs to which this noise is applied depends on the machine that we want to simulate. In **face matching** tasks, the probability p corresponds to the probability that both images in the pair represent the same person, while in

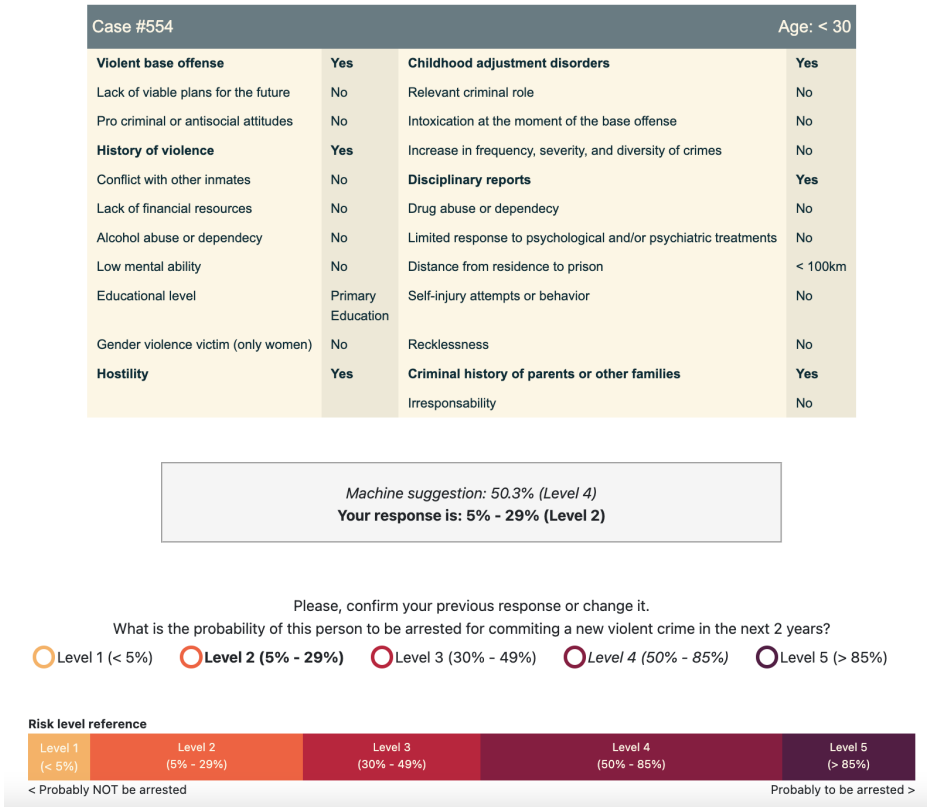


Figure 5.4: Interface for the suggestion of the machine with criminal recidivism tasks

criminal recidivism prediction tasks, p corresponds to the probability that the prisoner reoffends by committing a violent crime in the following two years. To carry out the experiments, we simulate two types of machines: static accuracy machines (experiments 1 and 2) and variable accuracy machines (experiment 3). These machines are described below.

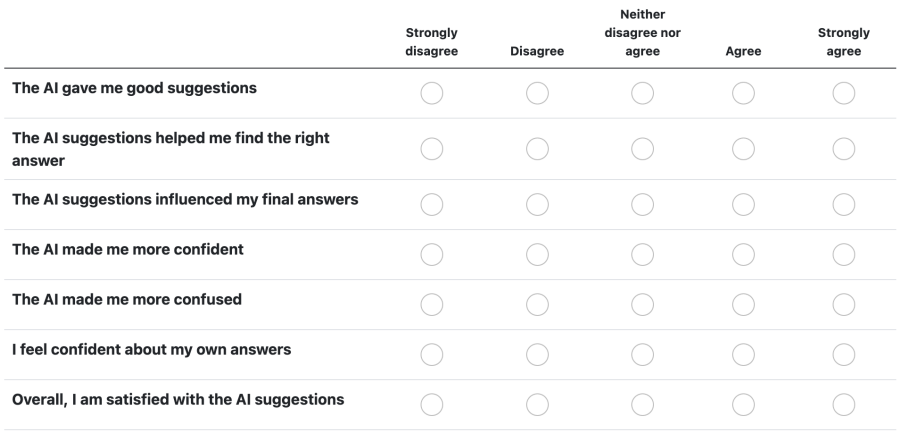


Figure 5.5: Exit survey screenshots for participants that received machine suggestions during the tasks.

Static machines Both for **face matching** and for **criminal recidivism prediction** tasks we simulate three machines with different accuracy: a high (95%) accuracy machine, a low (50%) accuracy machine, and a misleading machine (5% accuracy). We ensure that, in the criminal recidivism scenario, the AUC of the high-accuracy machine reflects that of RisCanvi (0.7 AUC), and similarly for that of the degraded-accuracy machines: the low accuracy machine has an AUC of 0.5, and the misleading machine has an AUC of 0.3. Each of these machines, both for face matching tasks and criminal recidivism tasks, is a realization of the probabilistic situations we simulate, which means that they have exactly the target accuracy (and AUC) in every experiment.

Variable machines We simulate three variable accuracy machines: 1) *INC machine* simulates a model that increases its accuracy over time, and is defined as the concatenation of three machines: misleading - low - high accuracy machines (each of the machines solving one third of the total number of tasks); 2) *DEC Machine* simulates

a model that decreases its accuracy over time, and is defined as the concatenation of the same machines as before, but in reverse order; and 3) *FAIL Machine* simulates a model that temporarily suffers an error, and is defined as the concatenation of three machines: high - misleading - high accuracy machines (each of them solving one third of the total number of tasks).

Regarding change notification in Experiment 3, there were two conditions: without notification, and with notification. Participants in the first condition were not told anything about variable machine accuracy. Participants in the notification condition, before starting the survey, were shown the following message:

> *There are three AIs, named machine A, machine B, and machine C. We will notify you every time there is a change.*

The notification message was:

> *Next, you will receive suggestions from machine {A / B / C}.*

5.3.3 Measurements

Participant accuracy We compute the fraction of correct responses, with respect to the ground truth, given by the participant before (*initial accuracy*) and after (*final accuracy*) seeing the machine suggestion. In the analysis we show the macro-average across participants of the initial and final accuracy.

Interaction Factors Given a task t solved by a participant, let r_i be the *participant's initial response*, r_f the *participant's final response*, and m the *machine suggestion*. In **face matching** tasks,

$r_i, r_f \in \{-2, -1, 1, 2\} \equiv \{$*No, Probably not, Probably yes, Yes*$\}$. In **criminal recidivism** tasks, $r_i, r_f \in \{1, 2, 3, 4, 5\} \equiv \{$*Level 1, Level 2, Level 3, Level 4, Level 5*$\}$. The *machine suggestion* m is the numerical representation of the probability given by the machine in the corresponding response interval. So, in face matching tasks, $m = -2 + 4p \in [-2, 2]$, being $p \in [0, 1]$ the similarity score given by the machine. In criminal recidivism tasks,

$$m = c + (p - a)\frac{(d - c)}{(b - a)} \in [0, 5]$$

being $p \in [0, 1]$ the probability of recidivism given by the machine,

$$[a, b] \in \{[0, 0.05], [0.05, 0.30], [0.30, 0.50], [0.50, 0.85], [0.85, 1.00]\}$$

the intervals defined by the levels of risk, and

$$[c, d] \in \{[0, 1], [1, 2], [2, 3], [3, 4], [4, 5]\}$$

the representation of the previous intervals in the response ranges. So, we define:

Influence Factor: We measured how much the machine suggestion influenced the participant's response. This value, that we denote by IF, ranges from -4 to 4. We define

$$IF(r_i, m, r_f) = \begin{cases} -|r_f - r_i| & \text{if } |m - r_i| < 1 \\ \frac{r_f - r_i}{m - r_i} & \text{if } |m - r_i| \geq 1 \end{cases}$$

So defined, the influence will be negative when there is a change between participant's initial and final response, and this change is in the opposite direction of what is recommended by the machine. Similarly, the influence will be positive when this change is in the direction what is recommended by the machine. In the analysis we show the macro-average of the influence factor. We also measure the

probability that this influence is positive, $P(IF > 0)$, zero $P(IF = 0)$, or negative $P(IF < 0)$.

Confirmation Probability: We measured the probability that the participant's initial response and the machine's suggestion match and the participant does not change their final response. This event, that we note as C, occurs when $m = r_i$ and $r_f = r_i$, that is, when the machine suggestion is the same as the participant's initial response and the participant's final response is the same as the initial one. In this case, the influence factor will be zero, not because there was no influence, but because the participant's response and machine suggestion coincided. In the analysis we show the probability that this occurs, that we note as $P(C)$.

5.4 Results

Table 5.2 summarizes our experimental results for Experiments 1 and 2, and Table 5.3 summarizes our experimental results for Experiment 3. We detail and explain these results next.

5.4.1 Experiment 1: With vs. Without DSS

In each scenario, for each set of different difficulty level, we compare two groups of participants: those who did not receive machine suggestions and those who received suggestions from the high accuracy machine. We compare their accuracy, their influence factor and their confirmation probability.

Easy Set In **face matching** tasks, almost all participants with no machine suggestions maintained their initial response when given the opportunity to modify it in the vast majority of tasks, making both

Face Matching								
Easy Set	a_i	a_f	δ	IF	$P(IF>0)$	$P(IF=0)$	$P(IF<0)$	$P(C)$
no machine	0.67 ±0.23	0.67 ±0.23	0	-	-	-	-	-
high-accuracy machine	0.71 ±0.20	0.80 ±0.20	+0.09	0.01	0.13	0.78	0.09	0.57
low-accuracy machine	0.69 ±0.21	0.66 ±0.21	-0.03	0.05	0.13	0.81	0.06	0.40
misleading machine	0.64 ±0.28	0.61 ±0.30	-0.03	-0.03	0.07	0.87	0.06	0.32
Hard Set	a_i	a_f	δ	IF	$P(IF>0)$	$P(IF=0)$	$P(IF<0)$	$P(C)$
no machine	0.57 ±0.20	0.57 ±0.20	0	-	-	-	-	-
high-accuracy machine	0.58 ±0.23	0.65 ±0.22	+0.07	0.13	0.15	0.81	0.03	0.34
low-accuracy machine	0.55 ±0.24	0.56 ±0.22	+0.01	0.24	0.23	0.75	0.02	0.34
misleading machine	0.54 ±0.22	0.38 ±0.24	-0.16	0.21	0.22	0.74	0.04	0.26
Criminal Recidivism								
Easy Set	a_i	a_f	δ	IF	$P(IF>0)$	$P(IF=0)$	$P(IF<0)$	$P(C)$
no machine	0.82 ±0.18	0.82 ±0.18	0	-	-	-	-	-
high-accuracy machine	0.83 ±0.16	0.88 ±0.17	+0.05	0.06	0.20	0.73	0.07	0.44
low-accuracy machine	0.79 ±0.19	0.79 ±0.18	0	0.10	0.25	0.67	0.07	0.37
misleading machine	0.71 ±0.16	0.58 ±0.17	-0.13	0.06	0.21	0.68	0.10	0.42
Hard Set	a_i	a_f	δ	IF	$P(IF>0)$	$P(IF=0)$	$P(IF<0)$	$P(C)$
no machine	0.54 ±0.20	0.54 ±0.20	0	-	-	-	-	-
high-accuracy machine	0.69 ±0.22	0.73 ±0.21	+0.04	0.07	0.25	0.65	0.10	0.34
low-accuracy machine	0.63 ±0.24	0.63 ±0.19	0	0.12	0.27	0.65	0.08	0.41
misleading machine	0.54 ±0.25	0.39 ±0.26	-0.15	-0.08	0.15	0.66	0.19	0.48

Table 5.2: Average participant initial accuracy a_i, final accuracy a_f, difference among both δ, influence factor IF, probability that this influence is positive $P(IF > 0)$, probability that this influence is neutral $P(IF = 0)$, probability that this influence is negative $P(IF < 0)$, and probability of confirmation $P(C)$ for Experiments 1 and 2 with the *Easy Set* and the *Hard Set*. There are 20 participants for every row.

the initial and final accuracy 0.67±0.23. Results from participants interacting with the high-accuracy machine suggest that the support of an accurate machine improves human performance (see Figures 5.6 and 5.10). Similarly, in **criminal recidivism prediction** tasks, almost all participants with no machine suggestions maintained their initial response, making both the initial and final accuracy 0.82±0.18. Results from participants interacting with the high-accuracy machine show that the support of an accurate machine significantly improves

Face Matching								
Easy Set	a_i	a_f	δ	IF	$P(IF>0)$	$P(IF=0)$	$P(IF<0)$	$P(C)$
INC (n)	0.64 ±0.21	0.64 ±0.22	0	0.00	0.11	0.82	0.07	0.45
INC (-)	0.64 ±0.21	0.63 ±0.18	-0.01	0.00	0.10	0.84	0.07	0.46
DEC (n)	0.70 ±0.21	0.69 ±0.21	-0.01	0.04	0.11	0.85	0.04	0.42
DEC (-)	0.71 ±0.18	0.67 ±0.21	-0.04	0.10	0.22	0.70	0.078	0.39
FAIL (n)	0.70 ±0.18	0.72 ±0.19	+0.02	-0.01	0.14	0.77	0.10	0.43
FAIL (-)	0.68 ±0.15	0.69 ±0.17	+0.01	0.00	0.12	0.79	0.09	0.47
Hard Set	a_i	a_f	δ	IF	$P(IF>0)$	$P(IF=0)$	$P(IF<0)$	$P(C)$
INC (n)	0.55 ±0.15	0.56 ±0.15	+0.01	0.18	0.18	0.78	0.04	0.35
INC (-)	0.49 ±0.17	0.49 ±0.17	0	0.20	0.22	0.73	0.05	0.33
DEC (n)	0.58 ±0.17	0.59 ±0.17	+0.01	0.11	0.15	0.80	0.05	0.36
DEC (-)	0.53 ±0.16	0.52 ±0.18	-0.01	0.13	0.19	0.72	0.09	0.30
FAIL (n)	0.54 ±0.16	0.55 ±0.19	+0.01	0.15	0.15	0.83	0.02	0.37
FAIL (-)	0.57 ±0.13	0.60 ±0.18	+0.03	0.17	0.18	0.80	0.02	0.35
Criminal Recidivism								
Easy Set	a_i	a_f	δ	IF	$P(IF>0)$	$P(IF=0)$	$P(IF<0)$	$P(C)$
INC (n)	0.70 ±0.22	0.68 ±0.24	-0.02	0.08	0.25	0.66	0.09	0.33
INC (-)	0.75 ±0.18	0.74 ±0.18	-0.01	0.07	0.26	0.65	0.09	0.33
DEC (n)	0.81 ±0.18	0.79 ±0.20	-0.02	0.04	0.18	0.72	0.09	0.42
DEC (-)	0.75 ±0.21	0.75 ±0.22	0	0.04	0.19	0.72	0.08	0.41
FAIL (n)	0.85 ±0.18	0.88 ±0.16	+0.03	0.04	0.18	0.75	0.07	0.44
FAIL (-)	0.81 ±0.19	0.82 ±0.17	+0.01	0.03	0.22	0.65	0.13	0.40
Hard Set	a_i	a_f	δ	IF	$P(IF>0)$	$P(IF=0)$	$P(IF<0)$	$P(C)$
INC (n)	0.48 ±0.26	0.52 ±0.26	+0.04	0.15	0.31	0.63	0.07	0.36
INC (-)	0.51 ±0.22	0.56 ±0.24	+0.05	0.07	0.28	0.60	0.11	0.39
DEC (n)	0.60 ±0.24	0.60 ±0.22	0	0.04	0.24	0.63	0.13	0.39
DEC (-)	0.60 ±0.21	0.60 ±0.20	0	0.09	0.25	0.64	0.10	0.41
FAIL (n)	0.63 ±0.23	0.64 ±0.24	+0.01	0.05	0.18	0.75	0.07	0.40
FAIL (-)	0.57 ±0.25	0.65 ±0.24	+0.08	0.13	0.31	0.59	0.09	0.35

Table 5.3: Average participant initial accuracy a_i, final accuracy a_f, difference among both δ, influence factor IF, probability that this influence is positive $P(IF > 0)$, probability that this influence is neutral $P(IF = 0)$, probability that this influence is negative $P(IF < 0)$, and probability of confirmation $P(C)$ for Experiment 3 with the *Easy Set* and the *Hard Set*. There are 20 participants for every row. (n) stands for *with notification*, (-) for *with no notification*.

human performance (see Figure 5.8).

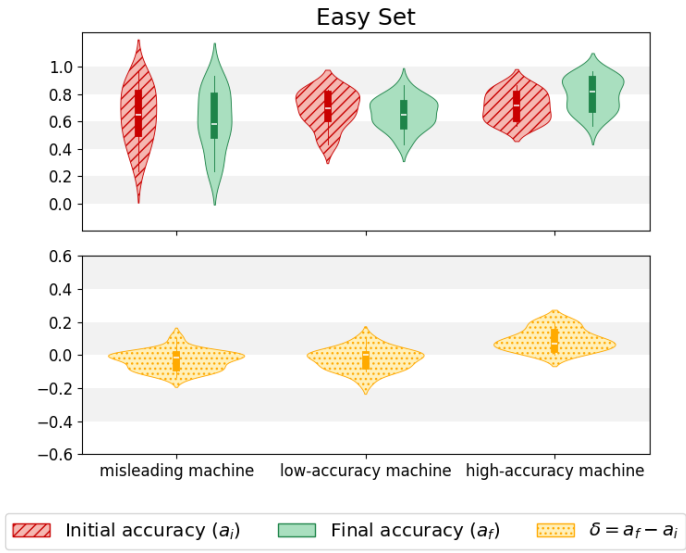


Figure 5.6: For **face matching** tasks, distributions of initial and final average accuracy across participants with the *Easy Set*, for misleading, low-accuracy, and high-accuracy machines in Experiments 1 and 2 with.

Hard Set In **face matching** tasks, almost all participants with no machine suggestions maintained their initial response when given the opportunity to modify it in the vast majority of tasks, making both the initial and final accuracy 0.57±0.20. As before, participants interacting with the high-accuracy machine got to improve their accuracy (see Figures 5.7 and 5.10). For these participants, the influence is higher than for those in the Easy Set. However, the results obtained in the exit survey show that participants perceive the high-accuracy machine more positively in the Easy Set than in the Hard Set, as shown in Figure 5.12. In **criminal recidivism prediction** tasks, almost all participants with no machine suggestions maintained their initial response, making both the initial and final accuracy 0.54±0.20. Results from participants interacting with the high-accuracy machine show that the support of an accurate machine improves human per-

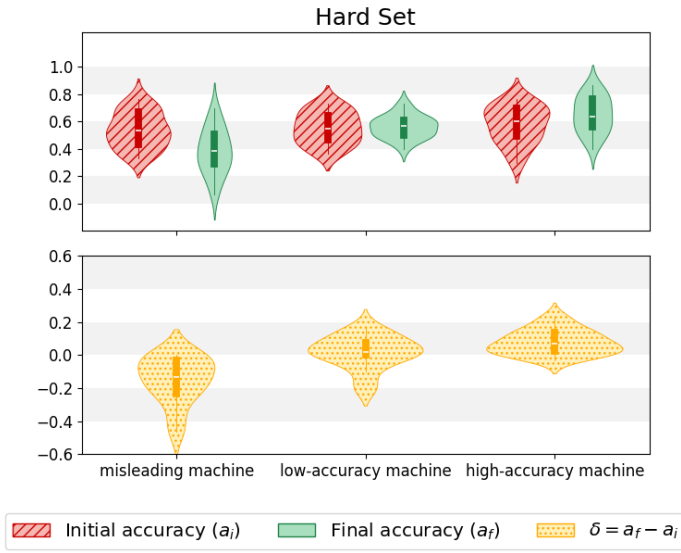


Figure 5.7: For **face matching** tasks, distributions of initial and final average accuracy across participants with the *Hard Set*, for misleading, low-accuracy, and high-accuracy machines in Experiments 1 and 2 with.

formance (see Figures 5.9 and 5.11). Although the influence of this machine in the Hard Set is similar to that of the Easy Set, the results of the exit survey (see Figure 5.13) show that participants perceive the high-accuracy machine significantly more positively in the Hard Set than in the Easy Set at $p < 0.001$.

5.4.2 Experiment 2: With Degraded DSS

In this experiment, we introduce the misleading machine and the low-accuracy machine.

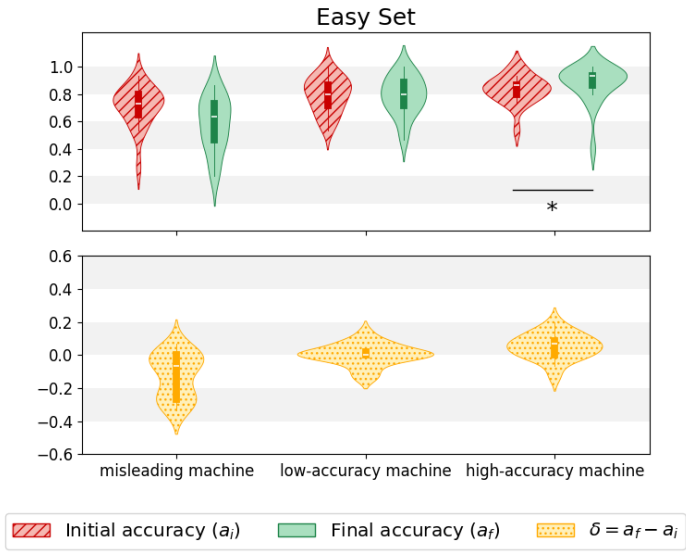


Figure 5.8: For **criminal recidivism** tasks, distributions of initial and final average accuracy across participants with the *Easy Set*, for misleading, low-accuracy, and high-accuracy machines in Experiments 1 and 2. The significance levels are labeled ($p < 0.05$: *).

Easy Set In **face matching** tasks, participants who interact with the low-accuracy machine and with the misleading machine experience a subtle drop in accuracy, suggesting that this type of support slightly deteriorates human performance (see Figures 5.6 and 5.10). Participants on a misleading machine showed a lower influence ($IF = -0.03$) compared to that of the high- and low-accuracy machines. This negative influence correlates with the perception that the participant has about the deteriorated accuracy machines. Results from the exit survey show that the participant perceives both the low-accuracy and the misleading machines significantly more negatively than the accurate machine at $p \ll 0.0001$. Also, the misleading machine is perceived significantly more negatively than the low-accuracy machine at $p \ll 0.0001$ (see left plot in Figure 5.12). In **criminal recidivism** tasks, participants who interact with the low-

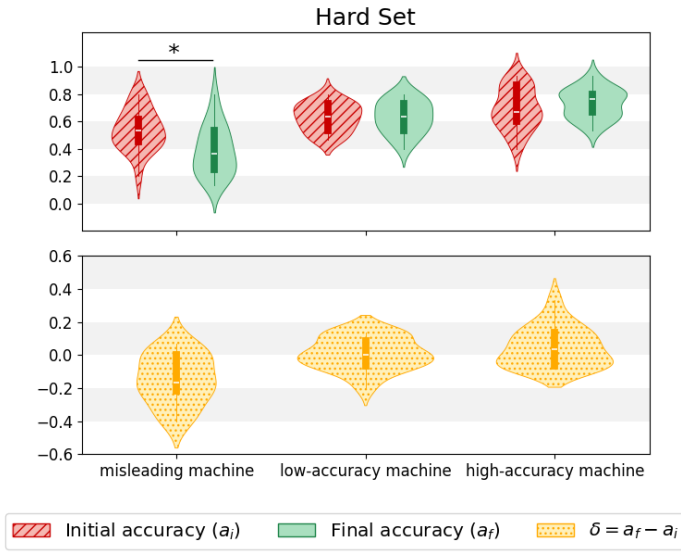


Figure 5.9: For **criminal recidivism** tasks, distributions of initial and final average accuracy across participants with the *Hard Set*, for misleading, low-accuracy, and high-accuracy machines in Experiments 1 and 2. The significance levels are labeled ($p < 0.05$: *).

accuracy machine do not experience a drop in accuracy, unlike the participants interacting with the misleading machine, whose decrease in accuracy is noticeable. This suggests that low accuracy support deteriorates human performance (see Figures 5.8 and 5.11). Unlike in face matching tasks, the influence factor of a low accuracy machine is similar to that of an accurate machine. This similarity in influence is correlated with the similar perceptions that the participant has about the deteriorated accuracy machines and the accurate machines (see left plot in Figure 5.13).

Hard Set In **face matching** tasks, for participants interacting with the low-accuracy machine, the participant's accuracy does not vary markedly (see Figures 5.7 and 5.10), and the influence is now no-

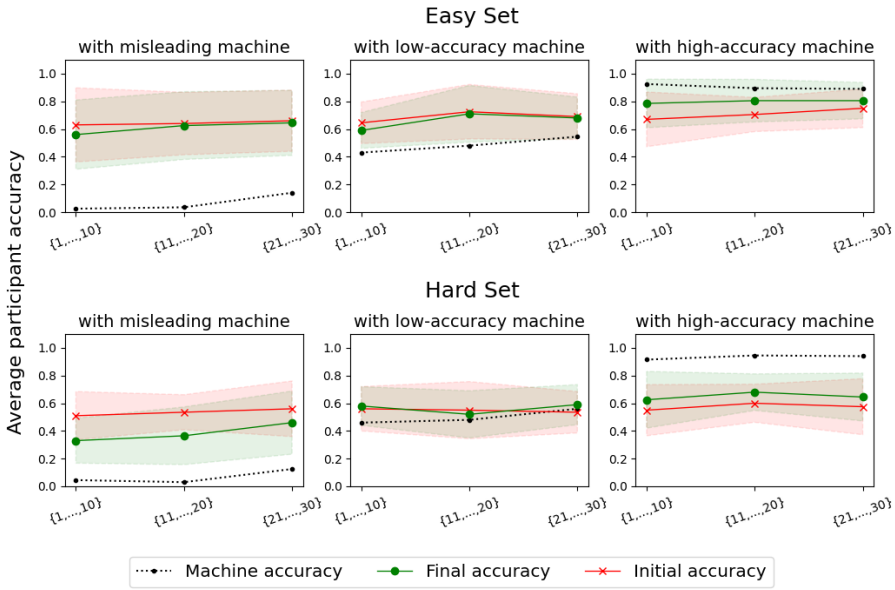


Figure 5.10: For **face matching** tasks, participant initial and final average accuracy with the *Easy Set* and the *Hard Set*, for misleading, low-accuracy, and high-accuracy machines, by set of tasks.

ticeable ($IF = 0.24$). For participants interacting with the misleading machine the influence is also high ($IF = 0.21$), but now there is a marked deterioration in accuracy (see Figures 5.7 and 5.10). These participants tend to be misled by the machine more often than those participants assigned to the same machine in the Easy Set, being their final accuracies significantly different at $p < 0.05$ even though their initial accuracies are not significantly different.

In Figure 5.12, we can see that for the participants who label pairs from the Easy Set, the more accurate the machine is, the more useful they find it. However, for the participants who label pairs from the Hard Set, the accuracy or inaccuracy of the machines does not affect as much the perceived usefulness of the machine. For the misleading and the low-accuracy machines, there are significant differences be-

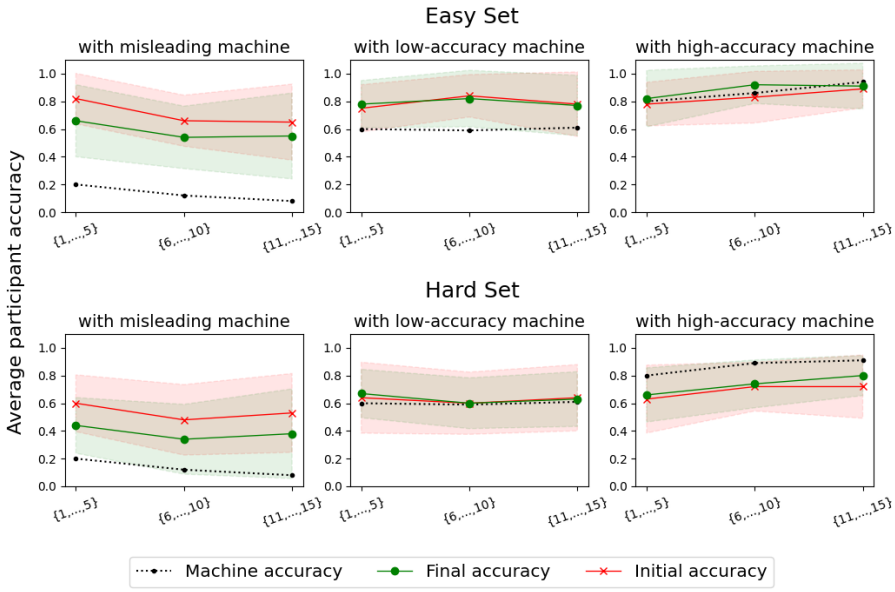


Figure 5.11: For **criminal recidivism** tasks, participant initial and final average accuracy with the *Easy Set* and the *Hard Set*, for misleading, low-accuracy, and high-accuracy machines, by set of tasks.

tween the distribution of perceived usefulness with the Easy Set and with the Hard Set ($p \ll 0.0001$ and $p \ll 0.0001$, respectively), with these machines being perceived as more useful for those participants who solved the task with the Hard Set. No significant differences are observed for the high-accuracy machine in different sets of difficulty. Also, as mentioned before, all three machines are perceived significantly different by those participants in the Easy Set, while no significant differences were found when comparing the same three machines in the Hard Set.

In **criminal recidivism** tasks, the accuracy of participants who interact with a low-accuracy machine does not vary (see Figures 5.9 and 5.11). For participants interacting with the misleading machine, there is a marked deterioration in accuracy (see Figures 5.7 and 5.10).

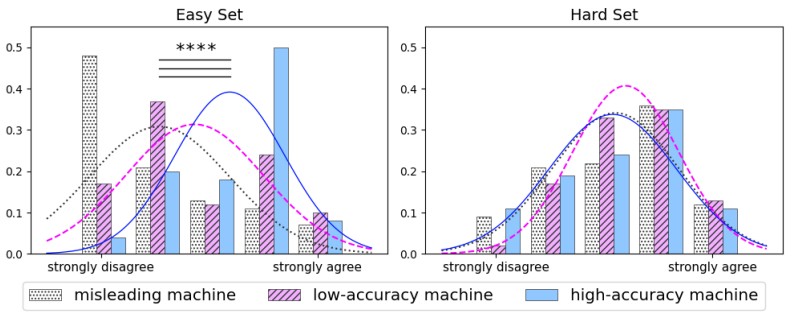


Figure 5.12: Results from the exit survey, from the participants who interacted with some static machine when solving **face matching** tasks. They were asked whether the AI 1. *gave them good suggestions* 2. *helped them find the right answer*, 3. *influenced their final answers*, 4. *made them more confident*, and whether they were 5. *satisfied with the AI suggestions*. We show the average of the responses across these four questions. The significance levels are labeled ($p < 0.0001$: ****).

These participants tend to be misled by the machine as those participants assigned to the same machine in the Easy Set. Their initial accuracies are significantly different ($p < 0.005$) across different levels of difficulty for participants assigned to deteriorated accuracy machines.

In Figure 5.13 we observe that for both participants in the Easy Set and participants in the Hard Set, the accuracy or inaccuracy of the machines does not affect the perceived usefulness of the machine, which suggests that the participant is not capable of distinguishing a deteriorated machine from an accurate machine.

The two-way ANOVA test (see Table 5.4) measuring the influence of task difficulty and machine accuracy over the difference between the initial accuracy and the final accuracy ($\delta = a_f - a_i$) reveals that the assigned machine (high accuracy machine, low accuracy machine, or misleading machine) has a significant influence on δ, with a large effect size η^2, in both applications (face matching and criminal recidi-

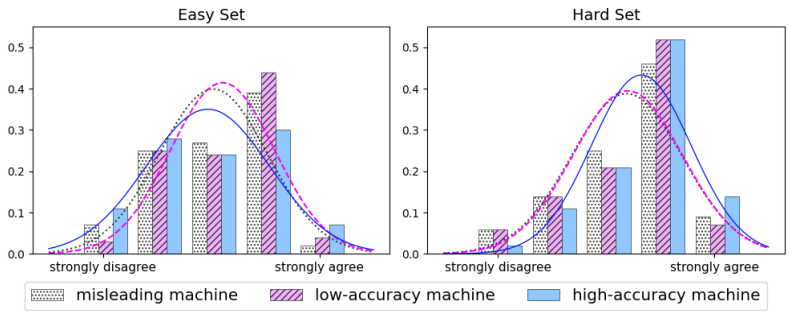


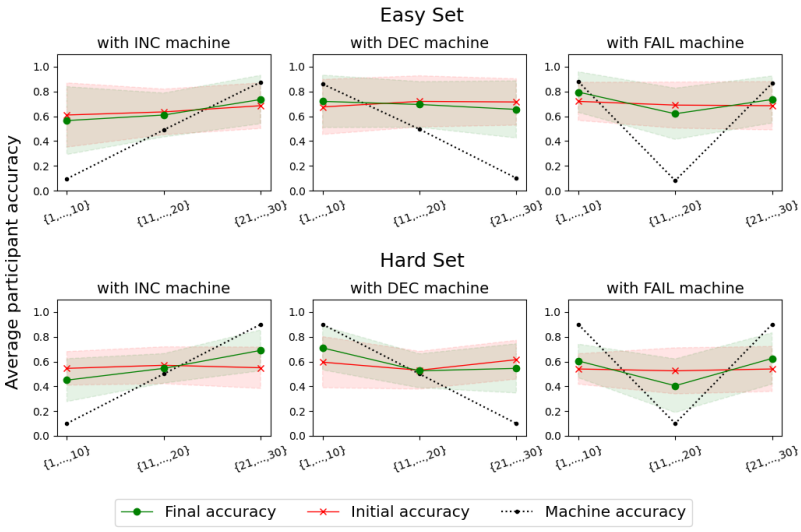
Figure 5.13: Results from the exit survey, from the participants who interacted with some static machine when solving **criminal recidivism prediction** tasks. They were asked whether the AI 1. *gave them good suggestions* 2. *helped them find the right answer*, 3. *influenced their final answers*, 4. *made them more confident*, and whether they were 5. *satisfied with the AI suggestions*. We show the average of the responses across these four questions.

vism). Additionally, in face matching tasks, the interaction between the assigned machine and the difficulty of the task is also significant.

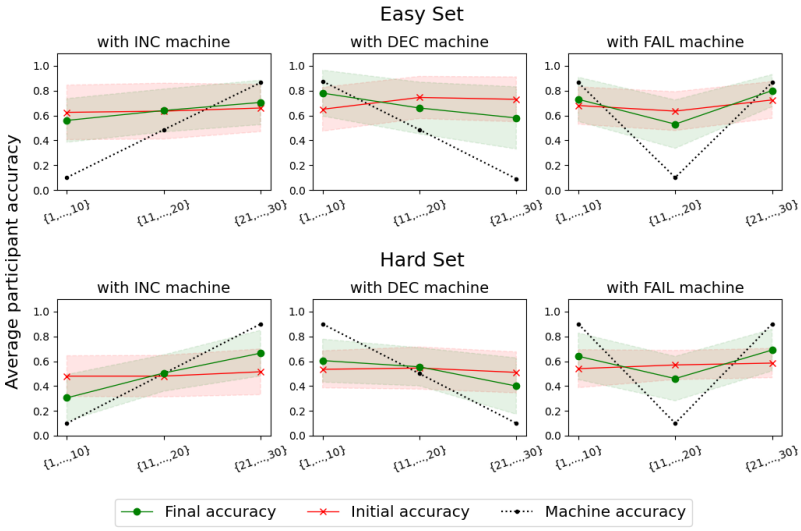
5.4.3 Experiment 3: With Variable DSS

For every set of pairs, we compare the group of participants who received suggestions from *INC Machine* (misleading - low - high), *DEC Machine* (high - low - misleading), and *FAIL Machine* (high - misleading - high). We distinguish between those participants who were notified every time the machine changed and those who were not.

Easy Set In **face matching** tasks, those participants assigned to the *INC Machine*, for both with (see Figure 5.14a) and without (see Figure 5.14b) notification, accuracy does not vary markedly, and the



(a) With notification



(b) Without notification

Figure 5.14: Participant accuracy with INC, DEC, and FAIL machines, with the *Easy Set* and the *Hard Set* in **face matching** tasks, with (a) and without (b) notification.

Face Matching						
Source	SS	DF	MS	F	p-value	η^2
Difficulty	0.027	1	0.027	3.52	0.06	0.03
Machine	0.607	2	0.303	39.54	****	**0.41**
Difficulty : Machine	0.129	2	0.064	8.39	***	**0.13**
Criminal Recidivism						
Source	SS	DF	MS	F	p-value	η^2
Difficulty	0.002	1	0.002	0.149	0.70	0.001
Machine	0.749	2	0.374	30.82	****	**0.35**
Difficulty : Machine	0.005	2	0.002	0.186	0.83	0.003

Table 5.4: Results from two-way ANOVA test in experiments with static machines (Experiments 1 and 2). The dependent variable is $\delta = a_f - a_i$ (the difference between the initial accuracy and the final accuracy). The significance levels are labeled ($p < 0.001$: ***; $p < 0.0001$: ****). "SS" stands for "Sum of Squares", "DF" for "Degrees of Freedom", "MS" for "Mean of Squares", "F" is the statistics, and η^2 indicates the size effect.

probability that there is no influence is above 0.80 in both cases. Those participants assigned to the *DEC Machine* with notification, accuracy does not vary markedly (see Figure 5.14a), and the influence is zero most of the times. For participants without notification, the results show a deterioration in the participant's accuracy, especially when the machine starts to deteriorate (see Figure 5.14b), and the positive influence is now noticeable. With the *FAIL Machine*, for both participants with and without notification, accuracy does not vary markedly, and the probability that the influence is zero is close to 0.80 in both cases.

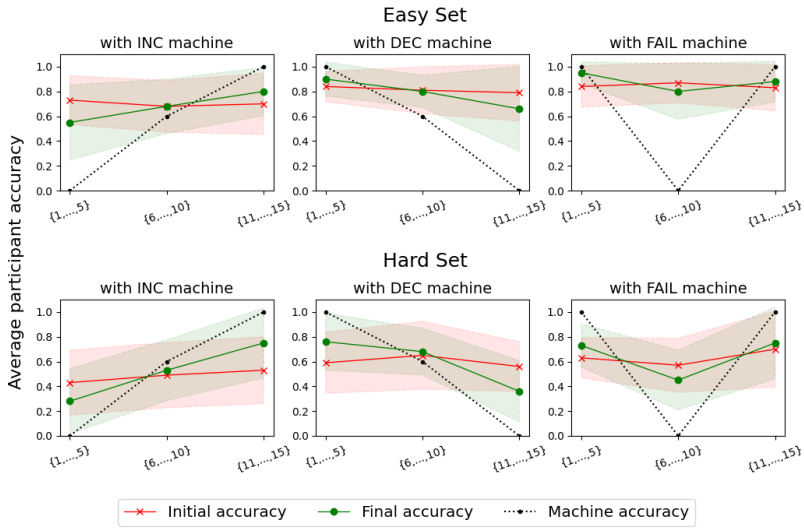
In **criminal recidivism** tasks those participants assigned to the *INC Machine*, for both with (see Figure 5.15a) and without (see Figure 5.15b) notification, show an improvement in accuracy, and the influence is positive at least 25% of the times. Those participants assigned

to the *DEC Machine* with (see Figure 5.15a) and without (see Figure 5.15b) notification, accuracy does not vary markedly, and the influence is zero more than 70% of the times. With the *FAIL Machine*, for both participants with and without notification, accuracy does not vary markedly, and the probability that the influence is zero is 0.75 and 0.65, respectively.

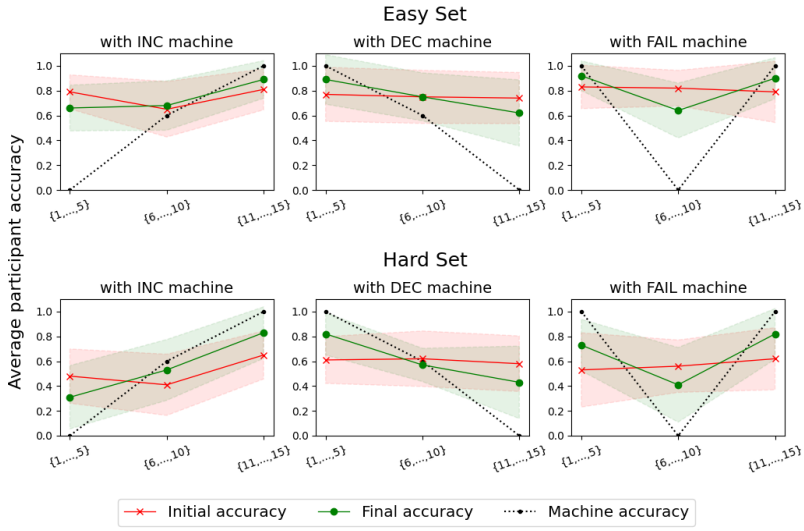
Hard Set In **face matching** tasks, with all *INC Machine*, *DEC Machine*, and *FAIL Machine*, for both participants with and without notification, accuracy does not vary markedly (see Figures 5.14a and 5.14b). We observe that those participants who interact with machines with no notification of change show higher values for the influence factor, and higher probabilities that the influence is positive, compared to those participants that receive the notification of change.

In **criminal recidivism** tasks, with both *INC Machine* and *DEC Machine*, for both participants with and without notification, accuracy does not vary markedly (see Figures 5.15a and 5.15b). With *FAIL Machine* without notification, the participant's accuracy improves slightly, and it does not vary markedly for those participants with notification.

For both face matching tasks (see Figure 5.14) and criminal recidivism tasks (see Figure 5.15), it is observed that in most cases the accuracy does not vary markedly regardless of whether the influence is high or low. This is because, as shown in the figures, participants are influenced by the variable accuracy machines both when they are accurate and when they are not.



(a) With notification



(b) Without notification

Figure 5.15: Participant accuracy with INC, DEC, and FAIL machines, with the *Easy Set* and the *Hard Set* in **criminal recidivism** tasks, with (a) and without (b) notification.

5.5 Discussion

A highly accurate machine may improve human performance, but the difficulty of the task might prevent the human from fully exploiting this advantage (**RQ1**).

For both face matching and criminal recidivism tasks, a high accuracy machine improves human performance, either in easy or hard tasks, which is aligned with previous works in the literature (Logg et al., 2019; Araujo et al., 2020). However, this improvement do not appear to stem from the participant's capacity to recognize the machine's high accuracy. In face matching tasks, for difficult pairs, the results obtained from the final questionnaire indicates that inaccurate machines are viewed as equally useful. In criminal recidivism tasks, either the cases are difficult or easy, the results obtained from the final questionnaire indicates that inaccurate machines are perceived as useful as the accurate machine. Thus, the difficulty experienced by the participant in solving a task seems to be correlated with their inability to recognize whether a support system is accurate or not.

High task difficulty allows an inaccurate machine to induce error more than an accurate machine can induce correctness, probably due to the participants' inability to really grasp how inaccurate the machine is (**RQ2**).

In face matching tasks, for degraded accuracy machines supporting human performance in easy tasks, the degradation of the participant's accuracy is hardly noticeable, suggesting that for a set of easy tasks the participant is able to solve without attending to the machine, as corroborated by the close-to-zero influence values. Observe that the minimal impact of the high-accuracy machine is partially due to the high confirmation rate (*i.e.*, the machine and user frequently agree, thus reducing the chance of influence), whereas the misleading machine exhibits a lower confirmation rate yet maintains a marginal influence (*i.e.*, in many tasks, the machine has opposed

the participant but did not alter their viewpoint). This suggests that for easy tasks, the participant knows how to solve it well, as it matches the high-accuracy machine and contradicts the misleading machine. These observations are supported by the ability of participants in the exit survey to distinguish between an accurate and an inaccurate machine. Not only is their perception correct, but this perception is also translated into practice, and participants are able to let themselves be guided by the machine only when it is accurate.

In contrast, in criminal recidivism tasks, a inaccurate machine can reduce the participant's accuracy even when the task is easy, probably because in these tasks the participant is no longer able to distinguish an accurate machine from an inaccurate one, as shown by the results of the exit survey. In fact, for these tasks, similarity is observed both when comparing the probabilities of positive influence by accurate and inaccurate machines. So, the misleading machine induces a participant to error that far exceeds the ability of the high-accuracy machine to induce correctness.

This also occurs in difficult face matching tasks. Our results suggest that in these tasks participants tend to be more influenced by low accuracy and misleading machines than by the highly accurate machine, while in easy tasks the opposite occurred. This same phenomenon also occurs, but more strongly, in difficult criminal recidivism tasks, where the decrease in accuracy for the participant assigned to an inaccurate machine is significant, while the improvement for the participant assigned to an accurate machine, although noticeable, is not significant.

Seen in this light, the results obtained for the criminal recidivism tasks appear to be an accentuation of the results obtained with the face matching tasks. This may be because the unfamiliarity participants experience with the recidivism tasks accentuates their difficulty, or, in other words, the familiarity participants experience with face matching tasks attenuates their difficulty. The results from the exit survey align with this idea. With familiar tasks (face matching), if

they are difficult, the participant is not able to perceive a difference in the usefulness of interacting with accurate or inaccurate machines, which is in line with some research in the literature, *e.g.*, Papenmeier et al. (2022). With unfamiliar tasks (criminal recidivism), this occurs regardless of whether they are difficult or easy, probably because the participant experiences the "easy tasks" with sufficient difficulty due to lack of experience.

Automation bias can be induced in a low-performing machine that initially provides accurate support. This can be mitigated with a simple notification that the machine has changed (**RQ3**).

For variable machines supporting human performance in face matching tasks that are easy, the machine influence is barely noticeable except in one case: when a machine initially functions with high precision but gradually loses accuracy. If the participant remains unaware of any change in the machine, they continues to rely on the machine, leading to mistakes. This aligns with the logic of some patterns observed in the literature. For instance, Dietvorst et al. (2015) observed that algorithmic aversion is seen to increase when the user sees that the machine fails. We observe something analogous: participants can move from algorithmic appreciation to automation bias after observing machine success. This situation can be prevented by notifying the participant about a machine change without revealing whether it is an improvement or a downgrade. This effect is not observed for variable machines that support human performance in difficult tasks, neither for variable machines in criminal recidivism tasks, probably because the participant is not able to clearly identify that the machine is performing accurately at the beginning, as the exit survey suggests.

It is notable to observe the distinction between outcomes from the low (50%) accuracy machine versus those from the respective machines with varying accuracy levels, increasing or decreasing. While all three machines maintain an average accuracy of 50%, they diverge in how their errors are distributed. Based on our findings, we

can deduce that for challenging tasks (difficult pairs in face matching tasks and both easy and difficult cases in criminal recidivism tasks), randomly distributed errors throughout the interaction promote error camouflage and thus increase the influence factor, compared to machines that accumulate errors at the beginning or at the end of the interaction flow, which have lower influence factors.

5.6 Conclusions

We noted that the difficulty of tasks shapes human-machine interactions in a variety of ways, even when problem-solving abilities are relatively similar. For instance, in face matching tasks, there is merely a 10% difference in average human accuracy between *easy* and *hard* tasks. Nevertheless, this small gap appears to be sufficient to notably change the influence of decision support. It is therefore crucial to understand that this challenge relates more to how the participant perceives the task than to their actual skill in solving it. Difficulty alone is not the only aspect of human perception that influences task resolution. The human's level of expertise in the task, or familiarity with it, appears to be another modulating factor in the relevance of a decision support system. This unfamiliarity, which could be seen as an added factor in how complex the task is perceived, is strongly related to difficulty: given a non-trivial task, the more unfamiliar it is, or the less experience one has solving it, the more difficult is perceived. Thus, high difficulty can affect the effectiveness of an accurate machine and can enhance the influence of an inaccurate machine. Conversely, low difficulty can enhance automation bias in the case of variable machines, more specifically those machines that start out accurate (thus eliciting appreciation) and later deteriorate. The combination of the task being perceived as sufficiently easy and the existence of an automated support system can promote fatigue, which would prevent the human from being alert to detect a decrease

in the support system's accuracy.

In this chapter we combined the interpretation of the influence factor with the confirmation factor, and that helped us to better understand patterns of interaction. However, the concept of influence allows for many approaches that need to be considered. We have noted that this influence can be negative (aversion), positive (appreciation/automation bias) or neutral (no influence). We observed a tendency towards no influence in easy tasks, and a more erratic tendency towards influence in the case of difficult tasks. This highlights the complexity of measuring influence, necessitating further study on when it is beneficial or detrimental for task completion. We also observed that participants' ability to distinguish accurate and inaccurate machines seems to be shaped by the distribution of the errors of the decision support system. A machine that makes errors evenly distributed throughout the interaction flow exerts more influence than one that accumulates its errors in some stretch of this flow.

However, the main factor that appears to override the ability to distinguish accurate from inaccurate machines is the perception of task complexity. Participants can distinguish between the accuracy and inaccuracy of decision support when the task is easy and familiar to them, but as soon as the task is difficult or unfamiliar enough, they lose this ability, making them more likely to be deceived. In the next chapter, we will explore ways to improve participants' ability to distinguish between accurate and inaccurate support systems, even under conditions of high difficulty and unfamiliarity with the tasks to be solved.

6 The Influence of Onboarding

This chapter is based on Estévez-Almenzar, M., Baeza-Yates, R., and Castillo, C. (2025). *Brief Onboarding Phase Improves Decision Support: Evidence from Two High-Risk Scenarios* (Submitted for publication)

6.1 Introduction

In a scenario in which a Decision Support System (DSS) provides decision support, we may highlight three main components: the task to be solved, the supporting system, and the person responsible for the final decision. In high-risk applications requiring meaningful human oversight, the human operator has the final say in the decision, and thus the role that the system plays depends on the mental model that the person has of the supporting system. In turn, the role of the human operator depends on the extent to which system recommendations are incorporated into human decisions. There is then a bidirectional influence between the human's mental model of the system, and its influence. This influence is modulated, among others, by the extent to which the human operator knows the task and can perform it accurately, i.e., task familiarity and difficulty.

In this chapter, we investigate whether an **onboarding phase**, i.e., a phase of preliminary contact, can affect this interaction. The incorporation of this phase is motivated by one of the main findings obtained in the experiments of the previous chapter: the difficulty of the task prevents humans from building a reliable idea about the machine with which they interact. Then, the key question for us in this chapter is whether this onboarding leads the human operator to develop a mental model of the capabilities and performance of a system that is better aligned with its actual accuracy.

The onboarding corresponds to a set of tasks in which human operators can see the system outcome and compare it with their own. We consider two variants of this, one in which human operators only see the comparison of labels, and one when they can additionally see the ground truth, and thus in case of discrepancies determine who was right.

We perform experiments in the two high-risk scenarios that we have been considering from the previous chapter. The first is face matching, in which a decision support system helps a person determine if two photos correspond to the same person or to different people. The use of remote biometrical AI-based tools has been declared as *high risk* in the AI Act by the European Commission (2024). The dangers and potential misuses of some of its applications have recently been studied (Lai and Rau, 2021; Raposo, 2024; Negri et al., 2024), and observed in the real world (The New York Times, 2020).

The second application is criminal recidivism prediction, in which a DSS advices on whether a prison inmate for which parole is being considered, will commit a new violent crime if released on parole. The use of AI-based tools for risk assessment in criminal justice has also been declared *high risk* in the AI Act by the European Commission (2024), and their potential negative effects have been investigated (Van Dijck, 2022; Scaria et al., 2024).

The differences in human familiarity with these scenarios arise natu-

rally: recognizing people's faces is something we do every day, while criminal risk prediction, for a non-expert human, is something new.

Based on the literature on DSS (outlined in §6.2) we study three interlinked research questions about onboarding a DSS.

RQ1 *How does an onboarding phase affect human-DSS interaction?*
We consider the effect of an onboarding phase prior to joint task-solving on human-DSS interaction. Previous work (§6.2) shows how human accuracy can be improved during this preliminary phase (by "training" the human operator), which suggests that preliminary phases are beneficial, so we expect to find similar findings in our experiments.

RQ2 *How does the presence of ground truth during the onboarding phase affect the human-DSS interaction?*
Most previous work (§6.2) presupposes the existence of ground truth and incorporates it during the preliminary phase. We want to explore the impact of not including the correct outcome on how the mental model of the system is constructed, as well as its alignment with the human's own confidence.

RQ3 *How does the interplay between task difficulty and the onboarding phase affect the human-DSS interaction?*
It is known that task difficulty influences how humans perceive DSS, but little is known about how it influences preliminary phases designed to sharpen this perception.

To study how difficulty in the task affects interactions, we use the two sets of tasks defined in the previous chapter (see *Task Selection* in §5.3).

We record all interactions with the system of a set of participants using a DSS in both tasks and both levels of difficulty. We use the same machines as in the previous chapter: one with a high (95%) accuracy, one with a low (50%) accuracy, and a misleading machine

(5% accuracy). We compare a control group (no onboarding) with a group that experiences an onboarding phase with visible or omitted ground truth. At the end, we use the same exit survey as in the previous chapter to determine perceptions of accuracy, beyond what is actually recorded as correct/incorrect decisions by participants. This protocol is described in detail (§6.3).

Our results (§6.4) indicate that an onboarding phase may improve human performance when assisted by a high-accuracy model and avoid reducing human performance with a low-accuracy or misleading model. However, the effectiveness of onboarding is affected by task difficulty and the presence of ground truth (§6.5). We conclude our paper with a summary of results, together with recommendations, limitations, and future work (§6.6).

6.2 Mental Models of the DSS

As defined by Johnson-Laird (1983), mental models are internal representations that people build based on their experiences in the real world. Norman (2014) state that mental models evolve as people integrate new observations into their reasoning. In the context of assisted decision-making, humans have been observed to adjust their mental models while working with a DSS in several studies by Kulesza et al. (2010); Lai et al. (2023); Ma et al. (2024). Moreover, Yin et al. (2019) observed that perceived accuracy has a significant effect on how often subjects follow the suggestions of the model, regardless of the actual model accuracy stated before starting the task. Indeed, in some cases, instead of explicitly stating the accuracy of the system, several researchers (Yu et al., 2016, 2017) observe that participants can perceive it, e.g., by offering instant feedback after interaction.

Previous works explore the impact of different workflow designs on the mental models that humans have in an assisted decision-making

scenario. Some designs include a pre-task training phase, during which users review the suggestion of the model and become familiar with the task (Chandrasekaran et al., 2018; Poursabzi-Sangdeh et al., 2021; Zhang et al., 2020). Even short preliminary phases included by Kulesza et al. (2012) have improved mental model soundness. Other studies such as Yin et al. (2019) propose including a feedback phase in the middle of the task. Most of these approaches assume that ground truth can be used during the preliminary phase, either training or feedback (Zhang et al., 2020; Yin et al., 2019; Chandrasekaran et al., 2018; Bansal et al., 2019a). However, other studies propose to work without showing the user the ground truth Poursabzi-Sangdeh et al. (2021). This can be relevant in settings in which feedback is delayed, such as predicting human behavior (e.g., criminal recidivism) with a time horizon measured in years (Holsinger et al., 2018). The impact of including or not including the ground truth of the task remains unclear.

In this work, we start from the idea of Carragher et al. (2024) that overcoming individual differences in task-solving is hard when it comes to complex, high-stakes tasks. Most of the surveyed research that uses a preliminary phase, such as Chandrasekaran et al. (2018); Poursabzi-Sangdeh et al. (2021); Zhang et al. (2020), adopts an approach based on a *training phase*, where users receive instant feedback, immediately after every single task, to show them the correct way of performing a task. In contrast, in this work we opt for an *onboarding phase*, in which feedback is not instantaneous, but instead given after a set of tasks, and the goal is not to train users but to make them aware of how the DSS recommendations may differ from their own. The goal of the onboarding phase is for users to develop a mental model of the DSS and to compare the recommendations of the DSS against their own decisions. The intention is to give users the opportunity to modulate their confidence in the support system, as well as their own confidence in solving the task, by comparing their own performance with that of the support system. Additionally, we

consider two factors that are still little explored in the literature: the presence or absence of ground truth during preliminary phases (Poursabzi-Sangdeh et al., 2021), and the effects of task difficulty on this phase (Salimzadeh et al., 2023).

6.3 Study Design

We conducted an online experiment similar to Experiments 1 and 2 in the previous chapter. Again, this new experiment was conducted twice: once with face matching tasks and once with criminal recidivism tasks. In this case, the experiment consisted of completing these tasks after the completion of an onboarding phase, which we describe later.

In previous experiments, the difficulty of the task proved to be one of the most influential factors in humans' ability to differentiate between an accurate and an inaccurate machine. Thus, in this new experiment, difficulty and machine accuracy remain part of the experimental conditions, along with a new condition: showing or hiding the ground truth during the onboarding phase. So, the following experimental conditions were combined:

1. *Task difficulty* (easy, hard), indicating the difficulty of the tasks assigned to the participant;

2. *Machine accuracy* (high, low, misleading), indicating the accuracy of the DSS working on its own;

3. *Ground truth* (present, absent), indicating whether ground truth is visible or not at the end of the onboarding.

6.3.1 Procedure

The experimental procedure of this experiment is similar to that of the experiments defined in the previous chapter except for the new onboarding phase. The experiment was conducted online as follows.

Participant Recruitment and Tasks

We recruited participants through Prolific. We considered four countries in continental Europe in which Prolific has large user bases: France, Germany, Italy, and Spain, plus the United Kingdom. We made sure that our sets of participants were gender balanced. Participants were paid 8.00 EUR per hour. To encourage participants' effort, we used a bonus payment: participants were informed (and reminded throughout the study) that if they managed to correctly solve more than 80% of the tasks they would receive an extra payment of 30%. The average completion time for each experiment was approximately 10 minutes. We recruited a total of 480 participants, 240 for each application (face matching and criminal recidivism).

We used our two sets of tasks defined for the experiments in the previous chapter (see Chapter 5, §5.3.1): the *Easy Set* and the *Hard Set*.

Onboarding phase

Participants had access to a DSS as described in §6.3.2 and underwent an onboarding phase. This phase consisted of evaluating a series of tasks, with no DSS and no feedback: 10 pairs of images for participants assigned to the face matching tasks (see Figure 6.3), and 5 profiles of people for participants assigned to the criminal recidivism tasks (see Figure 6.5). After completing these tasks, participants were shown a table containing their responses in one column, and

Figure 6.1: User interface for the final screen of the onboarding phase, in the face matching task (experiment 1). Those participants who went through the onboarding phase before starting the task evaluated 10 pairs of images with no instant machine suggestion. After solving these 10 tasks, they were shown a summary of their responses along with the machine's predictions for these tasks. Icons ✓ and ✗ were not shown in the ground truth omitted condition.

the machine suggestions in a second column. Those participants in the *visible ground truth* condition also saw the true response for every task in a third column. In addition, participants could view each of the tasks they had completed: the pair of images (see Figure 6.1) or the number corresponding to the profile of the incarcerated person, along with a drop-down menu to view the details of the case (see Figure 6.2).

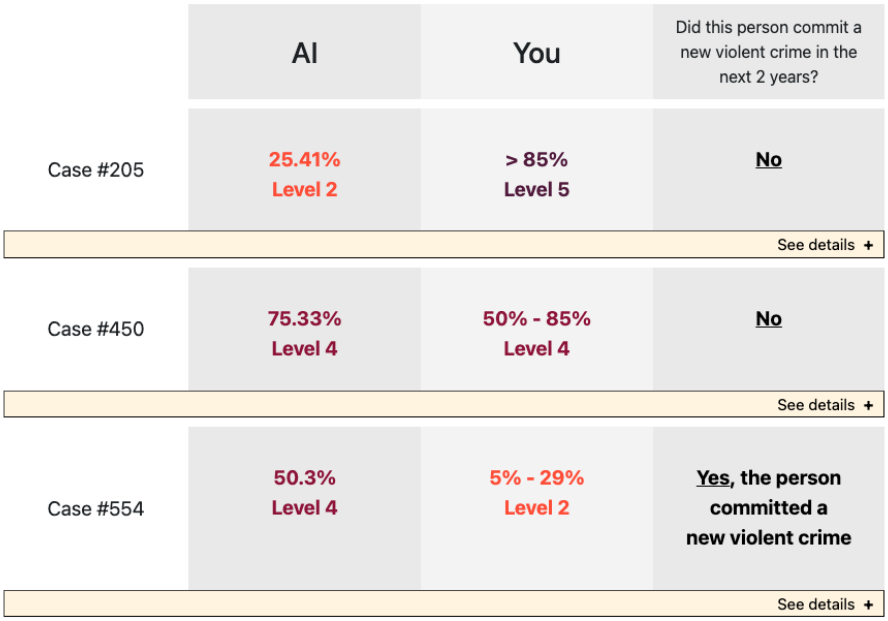


Figure 6.2: End of the onboarding phase for participants in Experiment 2 (criminal recidivism prediction). Those participants who went through the onboarding phase before starting the task evaluated 5 cases of imprisoned people with no instant machine suggestion. After solving these 5 task, at the end of the onboarding phase, they were shown a summary of their responses along with the machine's predictions for these tasks. Here you can see the interface for those participants in the visible ground truth condition. For those in the omitted ground truth condition, the column on the right was not shown.

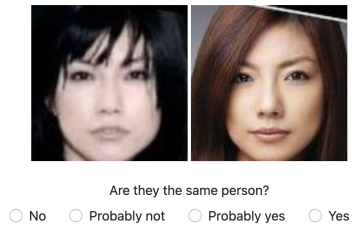


Figure 6.3: User interface for a face matching task.

Evaluation phase

The evaluation phase follows a structure practically identical to that described in the experiments in Chapter 5. We briefly recall it.

Participants evaluated one task at a time. Attention-check questions were included to filter out inattentive participants. Participants in face matching tasks were shown a pair of facial images, and they had to answer the question *Are they the same person?*, with the options *No*, *Probably not*, *Probably yes*, and *Yes*. The option "*Not sure*" was omitted to encourage them to make a decision in one direction or the other (see Figure 6.3). After answering the machine suggestion was shown together with the machine similarity score. Participants had the possibility to modify their answer (see Figure 6.4).

Participants in criminal recidivism tasks were shown one semi-synthetic profile representing a person who served a prison sentence at a time (see Figure 6.5), consisting of a list of 23 items that are used by RisCanvi to predict violent recidivism (Portela et al., 2024).

Participants had to answer the question *What is the probability of this person to be arrested for committing a new violent crime in the next 2 years?*, with the possible options *Level 1 (<5%)*, *Level 2 (5% - 29%)*, *Level 3 (30% - 49%)*, *Level 4 (50% - 85%)*, and *Level 5 (>85%)*.

Machine suggestion: "Yes", with similarity = 100.0%

Your response is "Probably yes"

Please, confirm your previous response or change it. Are they the same person?

No Probably not **Probably yes** *Yes*

Figure 6.4: User interface for the suggestion of the machine for a face matching task.

After answering, the machine suggestion was shown together with the recidivism probability estimated by the machine. Participants had the possibility to modify their answer (see Figure 6.6).

Participants in face matching tasks annotated 30 different pairs of images. In total, we collected 7,200 participant's annotations, and a total of 60 different pairs (30 easy, 30 hard) were annotated under various conditions. Participants in criminal recidivism tasks annotated 15 different cases of imprisoned people. In total, we collected 3,600 participant's annotations, and a total of 30 different cases (15 easy, 15 hard) were annotated under various conditions.

Exit survey

After completing the evaluation phase, the participants completed the exit survey described in Chapter 5 (see Figure 5.5). They were asked whether the AI: 1. *gave the participant good suggestions*, 2. *helped the participant find the right answer*, 3. *influenced the participant's final answers*, 4. *made the participant more confident*, and 5. *whether the participant was satisfied with the machine*, among other

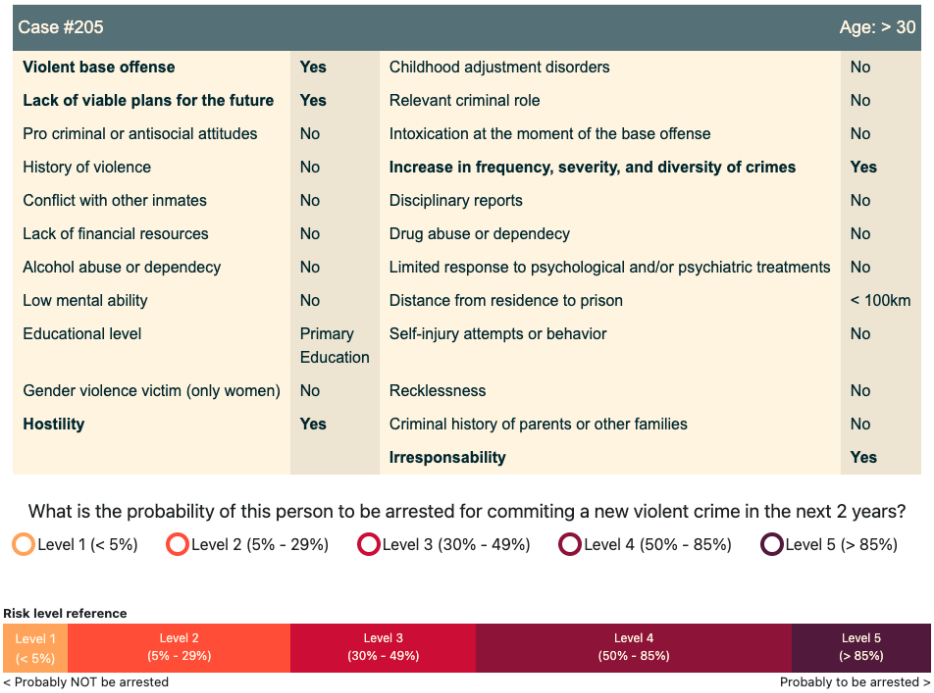


Figure 6.5: User interface for a prediction of criminal recidivism task. The profile presents the case of an imprisoned person. It consists of a list of 23 items that are used by RisCanvi to predict violent recidivism (Portela et al., 2024).

questions that were discarded for the analysis. For each question, a five-point Likert scale was used (*Strongly disagree*, *Disagree*, *Neither agree nor disagree*, *Agree*, *Strongly agree*).

6.3.2 Decision Support Accuracy

For this experiment, and for each application, we used the three static machines described in Chapter 5: one with *high accuracy*, one with *low accuracy* and one *misleading* machine (see §5.3.2).

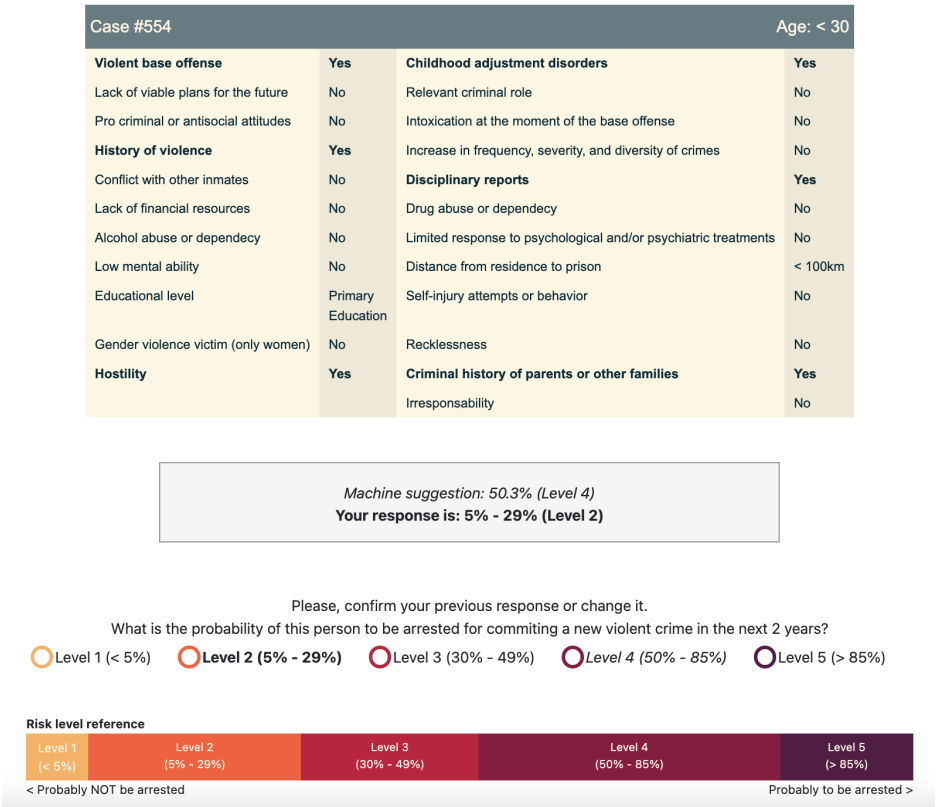


Figure 6.6: User interface for the suggestion of the machine with criminal recidivism tasks.

6.4 Results

Across all experimental groups, which are combinations of the experimental conditions described at the beginning of §6.3, we measured the *participant accuracy*, computed as the fraction of correct responses, with respect to the ground truth. As in previous experiments, we measured accuracy before receiving the machine suggestion (initial accuracy) and after receiving it (final accuracy).

Face Matching													
	no machine			misleading machine			low-acc. machine			high-acc. machine			
No onboarding	0.57	0.57	0	0.53	0.39	-0.14	0.55	0.56	+0.01	0.58	0.65	+0.07	Hard
Onboarding w/o ground truth				0.59	0.47	-0.12	0.60	0.56	-0.04	0.59	0.66	+0.07	Hard
Onboarding w/ ground truth				0.53	0.49	-0.04	0.56	0.53	-0.03	0.62	0.76	+0.14	Hard
No onboarding	0.67	0.67	0	0.64	0.61	-0.03	0.69	0.66	-0.03	0.70	0.80	+0.10	Easy
Onboarding w/o ground truth				0.69	0.56	-0.13	0.72	0.69	-0.03	0.85	0.89	+0.04	Easy
Onboarding w/ ground truth				0.74	0.72	-0.02	0.70	0.70	-0.00	0.78	0.82	+0.04	Easy
Criminal Recidivism													
	no machine			misleading machine			low-acc. machine			high-acc. machine			
No onboarding	0.54	0.54	0	0.54	0.39	-0.15	0.63	0.63	-0.00	0.69	0.73	+0.04	Hard
Onboarding w/o ground truth				0.39	0.26	-0.13	0.65	0.65	+0.00	0.72	0.82	+0.10	Hard
Onboarding w/ ground truth				0.43	0.35	-0.08	0.66	0.69	+0.03	0.67	0.73	+0.06	Hard
No onboarding	0.82	0.82	0	0.71	0.58	-0.13	0.79	0.79	-0.00	0.83	0.88	+0.05	Easy
Onboarding w/o ground truth				0.64	0.53	-0.11	0.79	0.76	-0.03	0.89	0.92	+0.03	Easy
Onboarding w/ ground truth				0.75	0.64	-0.11	0.83	0.81	-0.02	0.93	0.92	-0.01	Easy

Table 6.1: Average of the different accuracies of the participants, for each experimental condition. Each cell corresponds to a condition, and for each condition we show: initial final final − initial, where *initial* stands for the pre-assistance accuracy, and *final* stands for the post-assistance accuracy. In the row "No onboarding" we recall the accuracy of participant who did not go through the onboarding phase (those in the experiments 1 and 2 in Chapter 5). There are 20 participants in each condition.

In the following, we present the macro-average across participants of the initial and final accuracy. Table 6.1 summarizes our results, which we explain next.

In each scenario, for each set of different difficulty, we do two comparisons. First, we want to study the effect of including the onboarding phase, so we compare those participants who did not go through the onboarding phase and those who went through the onboarding phase in the *hidden ground-truth* condition. Second, we want to to study the effect of including the ground-truth during the onboarding phase,

so we compare and those participants who went through the onboarding phase in the *hidden ground-truth* condition with those who went through the onboarding phase in the *visible ground-truth* condition.

Easy Set

In **face matching** tasks, when participants solve the Easy Set, the onboarding phase seems to be unhelpful, even unfavorable. Those participants who received the machine suggestions without going through the onboarding phase improved their accuracy when interacting with the high-accuracy machine, while the final accuracy was worse for those interacting with the misleading machine (see Table 6.1 and Figure 6.7).

When including the onboarding phase, participants interacting with the high-accuracy machine become less influenced by its suggestions, both when ground truth is shown and when it is not, compared to those with no onboarding phase. Those participants interacting with the misleading machine are significantly more influenced by the machine when the onboarding phase is incorporated without ground truth at $p < 0.05$ (see Figure 6.8), which is undesirable, but this influence is almost nonexistent when the ground truth is included.

In both conditions (with and without onboarding) the results from participants with the low-accuracy machine are intermediate between the others.

In this case, the results from the exit survey show that the participants are capable of distinguishing whether the machine is accurate or not, being their perceptions significantly different (see Figure 6.9a), even when they did not go through the onboarding phase.

In **criminal recidivism** tasks, the onboarding phase seems to be unhelpful, but useful for the participant's ability to distinguish between an accurate and an inaccurate machine. Those participants

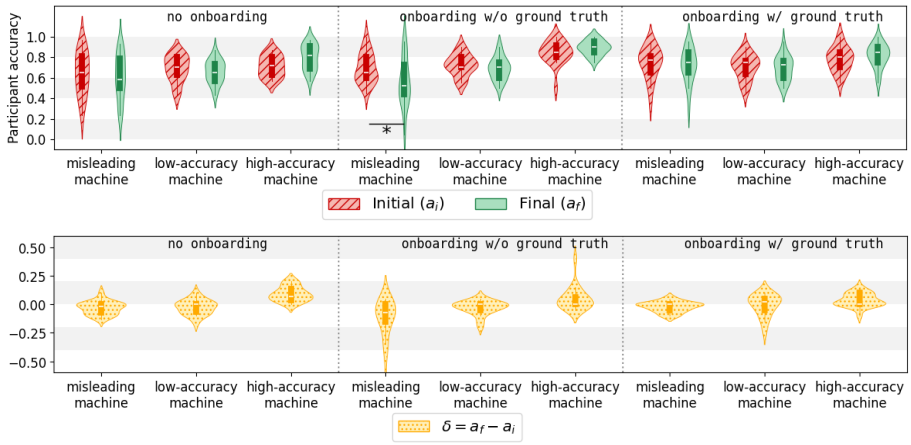


Figure 6.7: From participants in **face matching** tasks, initial and final average accuracy distributions for the **Easy Set**, for misleading, low-accuracy and high-accuracy machines. Three groups of participants are shown. From left to right: those who did not go through the onboarding phase, those who went through the onboarding phase with no ground truth, and those who went through the onboarding phase with ground truth. The significance levels are labeled ($p < 0.05$: *).

who received the machine suggestions without going through the onboarding phase improved their accuracy when interacting with the high-accuracy machine, making the difference between their initial and their final accuracy significant at $p < 0.01$, while the final accuracy was worse for those interacting with the misleading machine (see Figure 6.8). When including the onboarding phase, we do not observe any particular improvement regarding their final accuracy.

Nevertheless, by including the onboarding phase, the participants' capacity to differentiate between an accurate machine and an inaccurate machine improves significantly at $p < 0.001$ when the ground truth is not visible during the onboarding, and at $p < 0.0001$ when it is visible (see Figure 6.9b). This means that although the onboard-

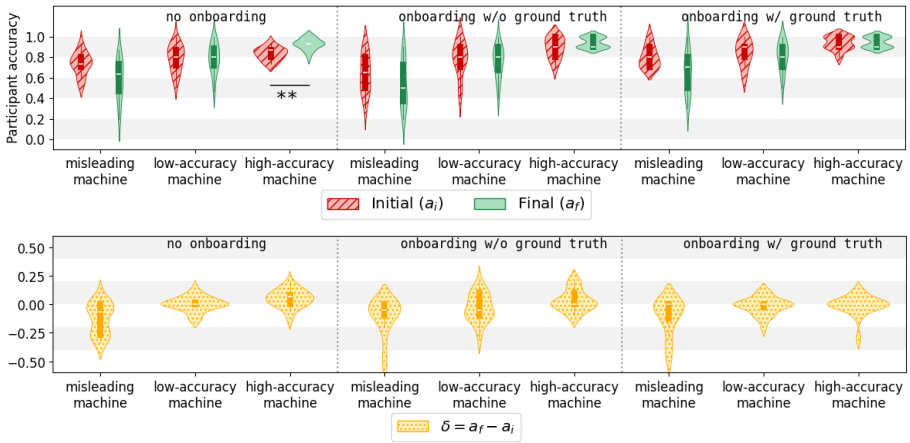
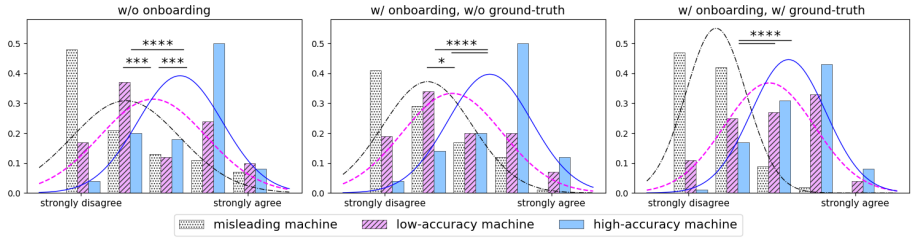


Figure 6.8: From participants in **criminal recidivism** tasks, initial and final average accuracy distributions for the **Easy Set**, for misleading, low-accuracy and high-accuracy machines. Three groups of participants are shown. From left to right: those who did not go through the onboarding phase, those who went through the onboarding phase with no ground truth, and those who went through the onboarding phase with ground truth. The significance levels are labeled ($p < 0.01$: **).

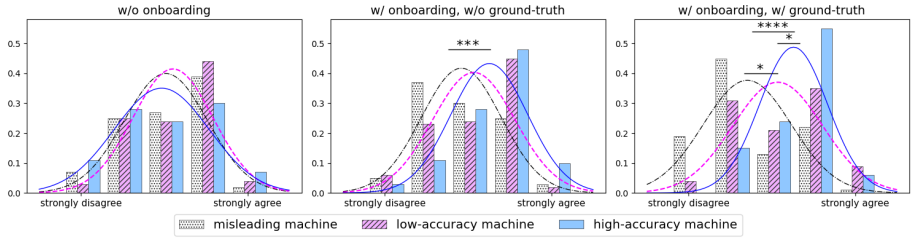
ing phase did not help participants improve their accuracy at a task that is relatively easy for them, it did have a significant impact on improving their perception of the machine.

Hard Set

In **face matching** tasks, those participants who received the machine suggestions without going through the onboarding phase improved their accuracy when interacting with the high-accuracy machine, while the final accuracy was worse than the initial one for those interacting with the misleading machine (see Table 6.1 and Figure 6.10).



(a) Face matching tasks



(b) Criminal recidivism tasks

Figure 6.9: Results from the exit survey from participants in face matching tasks (a) and criminal recidivism tasks (b), with the **Easy Set**. They were asked whether machine 1. *gave the participant good suggestions*, 2. *helped the participant find the right answer*, 3. *influenced the participant's final answers*, 4. *made the participant more confident*, and 5. *whether the participant was satisfied with the machine*. Participants had the possibility to answer *Strongly disagree / Disagree / Neither agree nor disagree / Agree / Strongly agree*. The significance levels are labeled ($p < 0.05$: *; $p < 0.001$: ***; $p < 0.0001$: ****).

Results suggest that with difficult tasks the onboarding phase helps to improve participant final accuracy. However, the presence of ground truth during the onboarding phase seems to be determinant to get a significant difference. Only when participants have seen the ground truth during the onboarding phase, they are capable to correctly adapt the influence that the machine exerts on them: the influence of the high-accuracy machine increases, making the final accuracy significantly different from the initial one ($p < 0.05$), while that of the misleading machine almost disappears, with respect to the influence that both machines exert on the participants who did not carry out the onboarding phase (see Figure 6.10). Again, in both conditions (with and without onboarding) the results from participants with the low-accuracy machine are intermediate between the others.

The results from the exit survey show that participants are, in general, unable to distinguish whether the machine is accurate or not. By including the onboarding phase, their capacity to differentiate an accurate machine from an inaccurate machine improves, and when this onboarding phase includes the ground truth, the improvement is significant at $p < 0.01$ (see Figure 6.12a).

In **criminal recidivism** tasks, those participants who received the machine suggestions without going through the onboarding phase improved their accuracy when interacting with the high-accuracy machine, while the final accuracy was significantly worse for those interacting with the misleading machine (see 6.1 and Figure 6.11).

When the onboarding phase is included, our results suggest that this phase also helps to improve participant final accuracy. The presence of the ground truth during the onboarding phase seems to be determinant to get a significant improvement with those participants interacting with the misleading machine. Only when participants have seen the ground truth during the onboarding phase, they are capable to correctly avoid being over-influenced by the inaccurate machine: their final accuracy stops being significantly different from their initial accuracy (see Figure 6.11). However, the presence of

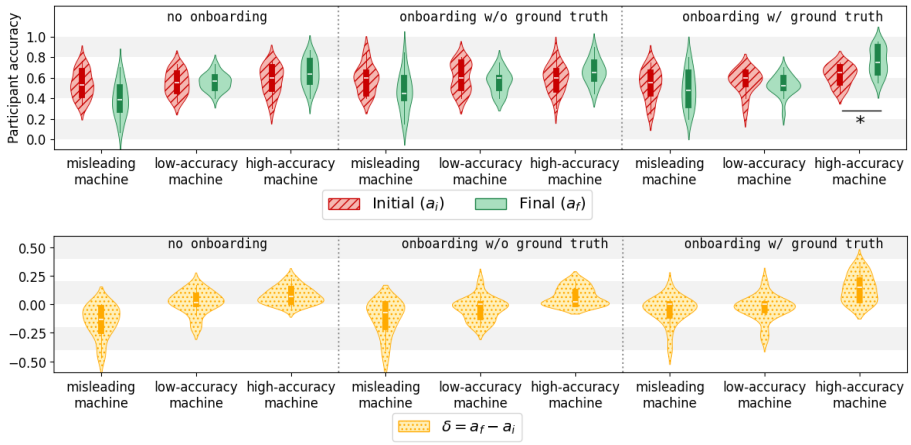


Figure 6.10: From participants in **face matching** tasks, initial and final average accuracy distributions for the **Hard Set**, for misleading, low-accuracy and high-accuracy machines. Three groups of participants are shown. From left to right: those who did not go through the onboarding phase, those who went through the onboarding phase with no ground truth, and those who went through the onboarding phase with ground truth. The significance levels are labeled ($p < 0.05$: *).

the ground truth during the onboarding phase seems not to be determinant to get a significant improvement with those participants interacting with the high-accuracy machine. When participants have gone through the onboarding phase with no ground truth, they are capable to correctly adapt the influence that the accurate machine exerts on them: the influence of the high-accuracy machine increases, and their final accuracy is significantly better than their initial accuracy at $p < 0.05$ (see Figure 6.11).

The results from the exit survey show that participants are not capable of distinguishing whether the machine is accurate or inaccurate either when they solves the task without going through the onboarding phase or when going through the onboarding phase (see Figure

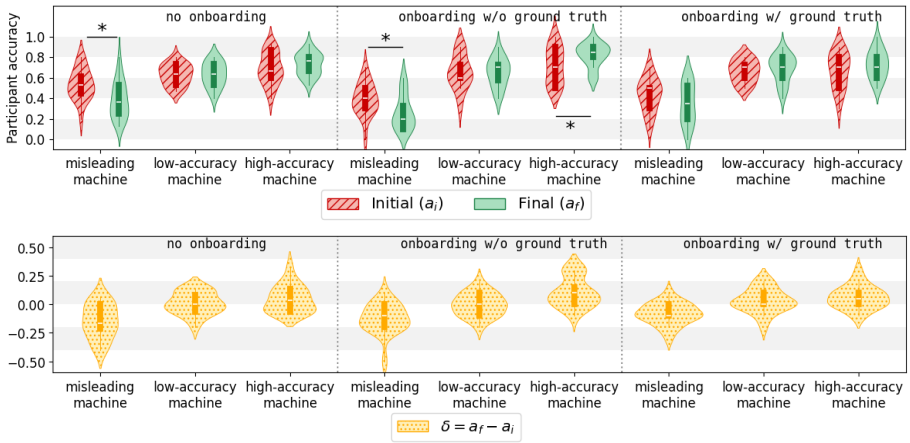
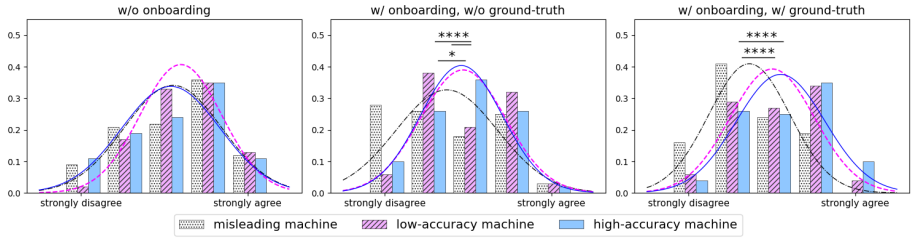


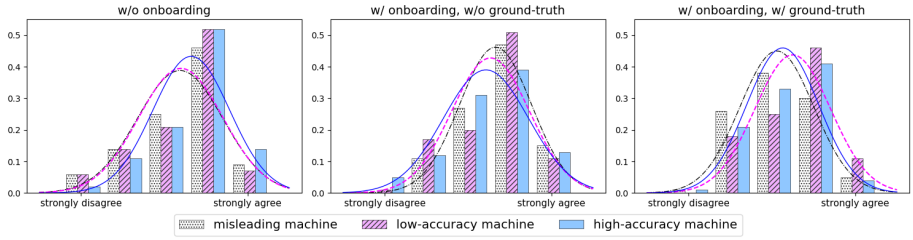
Figure 6.11: From participants in **criminal recidivism** tasks, initial and final average accuracy distributions for the **Hard Set**, for misleading, low-accuracy and high-accuracy machines. Three groups of participants are shown. From left to right: those who did not go through the onboarding phase, those who went through the onboarding phase with no ground truth, and those who went through the onboarding phase with ground truth. The significance levels are labeled ($p < 0.05$: *).

6.12b). This suggests that, for this scenario, including the onboarding phase is insufficient to make the participant distinguish between an accurate machine and an inaccurate machine (see Figure 6.12b).

Interestingly, for participants in **criminal recidivism** tasks we found noticeable differences between the **initial** accuracy of participants interacting with machines having different accuracies (see Figures 6.8 and 6.11). In general, it is expected that the machine will exert influence after showing the suggestion. However, we found some significant differences within pre-suggestion accuracies. In the Hard Set (see left plot in Figure 6.14), the initial accuracy of participants who went through onboarding without ground truth is significantly



(a) Face matching tasks



(b) Criminal recidivism tasks

Figure 6.12: Results from the exit survey from participants in face matching tasks (a) and in criminal recidivism tasks (b), with the **Hard Set**. They were asked whether machine 1. *gave the participant good suggestions*, 2. *helped the participant find the right answer*, 3. *influenced the participant's final answers*, 4. *made the participant more confident*, and 5. *whether the participant was satisfied with the machine*. Participants had the possibility to answer *Strongly disagree / Disagree / Neither agree nor disagree / Agree / Strongly agree*. The significance levels are labeled ($p < 0.05$: *; $p < 0.001$: ***; $p < 0.0001$: ****).

different between those who were assigned the misleading machine and those who were assigned the low-accuracy machine, and also between those who were assigned the misleading machine and those who were assigned the high-accuracy machine, at $p < 0.0005$. Similarly, the initial accuracy of participants who went through onboarding with ground truth is significantly different between those who were assigned the misleading machine and those who were assigned the high-accuracy machine at $p < 0.05$. Same significance is found when comparing the initial accuracy of those in the misleading machine and those in the low-accuracy machine. In the Easy Set (see right plot in Figure 6.14), the initial accuracy of participants who went through onboarding without ground truth is significantly different between those who were assigned the misleading machine and those who were assigned the high-accuracy machine at $p < 0.005$. Likewise, the initial accuracy of participants who went through onboarding with ground truth is significantly different between those who were assigned the misleading machine and those who were assigned the high-accuracy machine at $p < 0.001$. Additionally, the initial accuracy of participants who did not go through onboarding is significantly different between those participants who were assigned the misleading machine and those who were assigned the high-accuracy machine at $p < 0.05$.

These differences are not noticeable in **face matching** tasks (see Figure 6.13).

The three-way ANOVA test (see Table 6.2) measuring the influence of task difficulty, machine accuracy, and onboarding over the difference between the initial accuracy and the final accuracy ($\delta = a_f - a_i$) reveals that the assigned machine (high accuracy machine, low accuracy machine, or misleading machine) has a significant influence on δ, with a large effect size η^2, in both applications (face matching and criminal recidivism). Additionally, in face matching tasks, the interaction between (1) the assigned machine and the difficulty of the task, (2) the assigned machine and the presence of onboarding, and (3) the three independent variables (difficulty, machine, and

Face Matching						
Source	SS	DF	MS	F	p-value	η^2
Difficulty	0.014	1	0.014	1.306	0.254	0.006
Machine	1.388	2	0.694	67.17	****	**0.371**
Onboarding	0.033	1	0.033	3.161	0.077	0.014
Difficulty : Machine	0.064	2	0.032	3.114	*	**0.027**
Difficulty : Onboarding	0.014	1	0.014	1.306	0.254	0.006
Machine : Onboarding	0.077	2	0.038	3.715	*	**0.032**
Difficulty : Machine : Onboarding	0.064	2	0.032	3.114	*	**0.027**
Criminal Recidivism						
Source	SS	DF	MS	F	p-value	η^2
Difficulty	0.023	1	0.023	1.928	0.166	0.008
Machine	1.06	2	0.530	45.08	****	**0.283**
Onboarding	0.004	1	0.004	0.309	0.579	0.001
Difficulty : Machine	0.011	2	0.005	0.456	0.634	0.004
Difficulty : Onboarding	0.044	1	0.044	3.780	0.053	0.016
Machine : Onboarding	0.039	2	0.020	1.677	0.189	0.014
Difficulty : Machine : Onboarding	0.004	2	0.002	0.168	0.845	0.001

Table 6.2: Results from three-way ANOVA test. The dependent variable is $\delta = a_f - a_i$ (the difference between the initial accuracy and the final accuracy). The significance levels are labeled ($p < 0.05$: *; $p < 0.0001$: ****). "SS" stands for "Sum of Squares", "DF" for "Degrees of Freedom", "MS" for "Mean of Squares", "F" is the statistics, and η^2 indicates the size effect.

onboarding) are also significant, all of them with a small effect η^2.

6.5 Discussion

In both scenarios our results suggest that **onboarding improves accuracy and participants' ability to distinguish whether the machine is accurate or not** (RQ1). This aligns with some works in the literature (Kulesza et al., 2012; Do et al., 2024) showing that

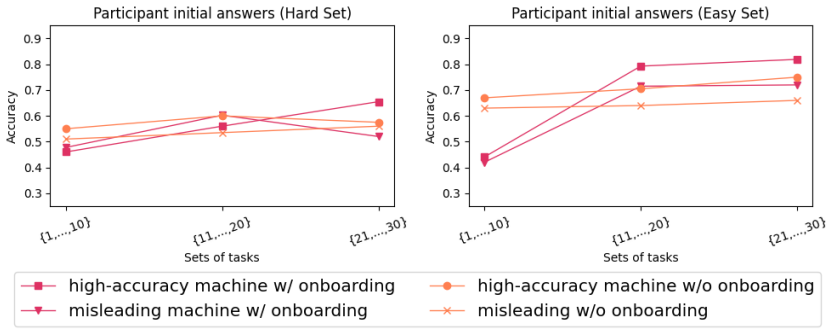


Figure 6.13: Evolution of the average initial (pre-assistance) accuracy of participants in **face matching** tasks, for different onboarding and machine conditions.

people's ability to correctly align their mental model of a DSS with its true performance improves after preliminary touch-base oriented phases.

However, the presence of **ground truth during this onboarding phase may be necessary to have a significant improvement** (RQ2). In face recognition tasks, the absence of ground truth can either lead to an accurate system not having a significant impact, or a less accurate system having more influence than desired. Our observations indicate that the former is prevalent in difficult tasks, while the latter occurs when tasks are easy. Furthermore, results from the exit survey indicate that participants can significantly differentiate between a precise and an imprecise machine only if they are exposed to the ground truth during the onboarding phase. In the absence of ground truth, while significant results were not achieved, there was a perceptible improvement compared to participants who skipped the onboarding phase. Interestingly, in the case of easy tasks, participants are capable of distinguishing an inaccurate machine even in the absence of the onboarding phase, but they are still at risk of being influenced by it. This aligns with previous work such as Poursabzi-Sangdeh et al. (2021), which shows that even when the

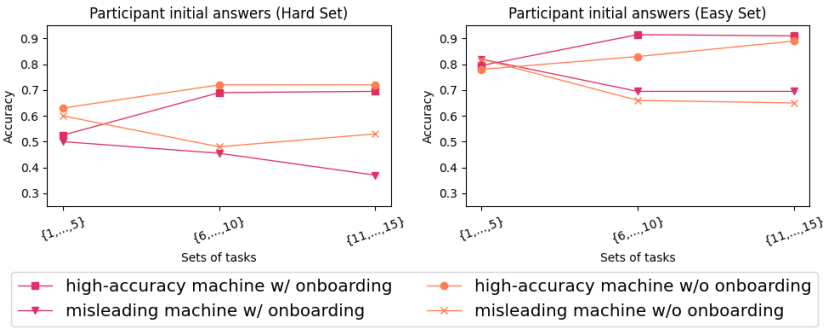


Figure 6.14: Evolution of the average initial (pre-assistance) accuracy of participants in **criminal recidivism** tasks, for different onboarding and machine conditions.

system's level of transparency helps people to detect when the model makes a mistake, the impact on people's willingness to distrust the model's prediction is not significant.

In the case of criminal recidivism prediction tasks, the presence of ground truth during the onboarding phase is not as decisive (RQ2). The absence of ground truth may result in the participant continuing to be significantly influenced by an inaccurate machine during the resolution of a difficult task (RQ3), which already happens to those participants who did not go through the onboarding phase. The presence of ground truth does not solve this problem, but it reduces it considerably. Results from the exit survey indicate that participants can significantly differentiate between a precise and an imprecise machine after going through the onboarding phase, either with or without ground truth, but only after solving easy tasks. **For difficult tasks, the onboarding phase appears to be insufficient for the participant to distinguish between imprecise and precise machines** (RQ3). This difference in behavior, which appears to depend on task difficulty, marks a difference between our work and those in the literature (Yu et al., 2016, 2017) that show that people are able to perceive how precise a model is without the

need to explicitly communicate the system's accuracy.

We observe an unexpected influence of decision support on participants. In criminal recidivism prediction tasks, we observe that the participants' initial answers (those given before seeing the machine suggestion) have different evolution over the workflow depending on the machine with which the human is interacting (see Figure 6.14).

Participants in the same onboarding phase condition (with or without) had similar initial accuracy to the first five tasks (see Figure 6.13). However, as the evaluation phase progresses, the initial accuracy of those participants interacting with an accurate machine begins to be significantly better than the initial accuracy of those participants interacting with an inaccurate machine. This suggests that the machine impacts not only the participant's final response (the one given after seeing the machine's suggestion, which would be expected), but also the response that the participant gives on their own, before getting the machine's suggestion. Interestingly, this occurs only with the criminal recidivism prediction tasks. In the case of face recognition tasks, this difference is not observed. It is plausible that this is related to the difference in familiarity or prior knowledge that the participant has between these two scenarios: while the ability to recognize and identify faces is something already learned, even everyday, the task of predicting a crime requires more specific experience and even specialist knowledge. This observation is aligned with some results in the field of Explainable AI (Wang and Yin, 2021) where it has been observed that the effectiveness of explanations noticeably depends on the varying levels of domain expertise.

6.6 Conclusions

This study investigated the impact of a brief onboarding phase on human interaction with Decision Support Systems (DSS) in two

high-risk scenarios: face matching and criminal recidivism prediction. Across both domains, incorporating a brief onboarding period was advantageous in aligning user perceptions with the system's actual performance. Our findings provide strong evidence that **even a brief onboarding can significantly influence whether users perceive system accuracy correctly** – especially when accompanied by ground truth – and thus help them decide to what extent to follow its recommendations. The onboarding phase demonstrated some advantages in improving task performance when the DSS is highly accurate, and in not reducing task performance when the DSS is inaccurate.

Experimental results also suggest that **onboarding's effectiveness depends on task difficulty and ground truth availability.** In hard settings, the presence of ground truth during onboarding was particularly important: it allowed users to correctly identify and weigh the reliability of machine suggestions. This was consistently observed in both applications. In facial matching tasks, the onboarding helps users to increase their trust in high-accuracy systems, while in criminal recidivism tasks, the onboarding helps users to decrease their trust in low-accuracy systems. Conversely, when the task was hard and there was no ground truth, onboarding was less useful. In easy task settings, the benefits of onboarding were milder. Participants generally performed well without assistance, and onboarding occasionally introduced confusion, or even increased erroneous influence from low-accuracy systems when ground truth was not included. This highlights a potential risk: onboarding may inadvertently lend credibility to poor decision aids unless supported by performance feedback mechanisms like ground truth exposure. In general, if the task is so easy that a DSS is not needed, we would recommend not using it.

An unforeseen insight emerging from our study concerns the **implicit influence of DSS interaction in how users learn to perform a task.** In the criminal recidivism scenario, we observed that exposure

to accurate systems led to a progressive increase in participants' non-assisted accuracy over time. This suggests a learning effect influenced by machine interaction. Such an effect was not evident in the face recognition task, likely due to the participants' familiarity with face matching and lack of familiarity with recidivism risk prediction. This suggests that onboarding benefits regarding learning are particularly significant in unfamiliar domains.

In both experiments, exit survey data reinforced these behavioral findings. In general, participants who underwent onboarding with visible ground truth were significantly **better at judging the accuracy of the DSS.** Task-specific differences can be observed here. For difficult tasks in the face matching scenario without ground truth, participants continued to struggle to differentiate between accurate and inaccurate systems, even after onboarding. However, the absence of ground truth for easy tasks in the criminal recidivism scenario is not decisive: participants are capable of differentiating between accurate and inaccurate systems even when they receive no feedback about the correctness of the system during onboarding.

7 Conclusions

7.1 Summary of Findings

The work presented in this dissertation contributes to the human-computer interaction in decision making literature by providing several findings summarized hereafter.

7.1.1 The Relevance of Human-Machine Complementarities

In Chapter 4, we investigated the complementary dynamics between human and machine performance in face recognition tasks, with a particular focus on the nature and alignment of their respective errors. Our goal was to understand when humans and facial recognition models make similar mistakes, when their errors diverge, and how this information can inform more efficient and ethically responsible hybrid decision-making strategies.

Our findings confirm that humans and machines exhibit distinct patterns of error, and that these differences can be strategically leveraged. First, we observed a high level of consistency in both human annotations and model outputs, allowing us to identify non-overlapping

errors, particularly those that machines are prone to (*e.g.*, false positives) but humans rarely make. Notably, humans outperform machines in detecting negative face pairs, which suggests that human oversight is especially valuable for preventing these types of erroneous identifications.

Second, we found that shared errors between multiple models often correspond to increased difficulty for human annotators, especially for false negatives. However, when only one model errs, humans are more likely to correctly classify the pair, highlighting the benefit of diversity in algorithmic decision-making and the importance of selecting the right cases for human review.

Third, the algorithm similarity score proved to be a useful predictor of potential error: there is a clear gap between similarity scores for correct and incorrect classifications, which can be used to flag cases for manual review. This scoring insight, combined with the observed human strengths, enabled us to design a targeted oversight strategy. By having humans review only a small subset (10%) of machine predictions, we were able to improve system accuracy by approximately 3 percentage points, with minimal additional annotation cost.

In practical terms, this means that a well-designed hybrid system can achieve higher accuracy and reliability than either humans or machines alone, while also minimizing resource expenditure. Our work underscores the necessity of human oversight, particularly in sensitive or high-stakes scenarios where false positives may have severe ethical or legal consequences.

Moreover, our analysis points to the influence of perceived gender expression and ethnic appearance in human decision-making, especially in correctly classifying machine false positives. This opens important ethical questions about implicit human biases and the representational biases embedded in training datasets, both of which need to be carefully considered when deploying hybrid systems in real-world applications.

In conclusion, in Chapter 4 we show that human and machine errors are neither redundant nor exclusive but complementary, and that these complementarities can be systematically exploited to enhance performance and accountability. The findings also emphasize that full automation in face recognition is neither technically mature nor ethically advisable at this stage. Ongoing human involvement will remain essential for safe and effective deployment, especially in light of emerging regulatory frameworks such as the European AI Act, which recognizes facial recognition as a high-risk application domain.

7.1.2 Task Difficulty and Familiarity Shaping Human Mental Models

In Chapter 5, we examined how task difficulty and the accuracy of decision support systems (DSS) influence human decision-making in high-risk domains. By introducing human-machine interaction to both face matching and criminal recidivism prediction tasks, we investigated whether and how human performance is affected by varying levels of machine accuracy: high and low accuracy, and misleading, variable machines. We also investigated how this relationship is modulated by perceived task complexity.

First, we found that accurate DSSs can improve human performance, but this improvement is significantly attenuated by task difficulty. Participants often failed to recognize when a machine was highly accurate, especially in difficult tasks. In both face matching and recidivism prediction, the perceived complexity of the task impaired users' ability to distinguish between accurate and inaccurate systems, undermining potential benefits of accurate decision support.

Second, we observed that inaccurate and misleading DSSs can degrade human performance, particularly in complex tasks or unfamiliar domains. In face matching tasks, participants were generally able to resist inaccurate support when tasks were easy, showing low influ-

ence from bad suggestions. However, in criminal recidivism prediction, even easy cases lead participants to rely on inaccurate support just as much as on accurate assistance. This suggests that familiarity with the task domain plays a critical role in moderating the influence of DSSs.

Third, we explored the impact of variable accuracy machines. Systems that start with high performance and later degrade pose a unique risk: they induce automation bias, where early trust in the system persists even after its performance drops. Notably, simple notifications informing users of a change in the system's accuracy, without specifying the direction of change, can mitigate this risk. This finding suggest that simple notification of changes may allow users to adjust their reliance on the DSS. This effect was especially evident in easy face matching tasks, but not in complex or unfamiliar tasks, where users struggled to evaluate performance regardless of changes.

Furthermore, our findings reveal that how errors are distributed across a sequence of interactions significantly shapes influence. Systems with randomly distributed errors were more misleading than those whose errors were either accumulated at the beginning or at the end of the interaction.

In conclusion, Chapter 5 demonstrates that task difficulty (as perceived by users) is a powerful modulator of human-machine interaction outcomes. Increases in perceived difficulty reduce the human's ability to evaluate machine accuracy, making them more susceptible to both automation bias and consequently misleading support.

7.1.3 Onboarding Mitigates the Consequences of Task Difficulty and Unfamiliarity

In Chapter 6, we investigated how an onboarding phase (a brief period of preliminary contact with the task to be solved and the DSS) affects the dynamics of human-machine collaboration in two high-risk decision-making scenarios: face matching and criminal recidivism prediction. We explored whether onboarding helps users develop an accurate mental model of the DSS, particularly under varying task difficulties and with or without access to ground truth during the onboarding, which is relevant in settings in which feedback is delayed, such as predicting human behavior (e.g., criminal recidivism) with a time horizon measured in years.

Our findings provide evidence that onboarding enhances users' ability to assess the reliability of the DSS, especially when the system is accurate and when the onboarding phase includes ground truth. This responds to our first research question: onboarding improves both user performance and system trust calibration, helping users decide when to rely on or reject machine suggestions.

Crucially, the presence of ground truth during onboarding significantly amplifies this effect. In face matching tasks, ground truth made the difference between users gaining trust in an accurate system or being misled by an inaccurate one. Without ground truth, participants often failed to discern the system's reliability, especially in difficult tasks. In contrast, for criminal recidivism prediction, a less familiar task, participants also benefited from onboarding, but the presence of ground truth was less decisive. Nevertheless, it still contributed to reducing the influence of low-accuracy systems, especially in easier cases.

As we have observed throughout this work, task difficulty played a central role in shaping the effectiveness of onboarding. When tasks were difficult and unfamiliar, onboarding alone (without ground

truth) was insufficient to help users distinguish between high and low accuracy systems. This limitation underscores the importance of both task familiarity and feedback mechanisms in ensuring that preliminary phases leads to informed decision-making.

An important and unexpected insight was the observation that interaction with the DSS affected how participants performed even before viewing its suggestions, particularly in the criminal recidivism task. Users exposed to high-accuracy systems gradually improved their unaided judgments over time, suggesting that machine interaction itself may have a learning effect in unfamiliar domains. This effect was not observed in the face matching task, likely due to the participants' pre-existing competence in that domain.

Overall, our results emphasize that a brief onboarding phase can build user trust where warranted and reduce overreliance on faulty systems. However, its effectiveness depends strongly on both the cognitive complexity of the task and the presence of performance feedback. In tasks where users already perform well unaided, onboarding without ground truth may even introduce noise or inadvertently legitimize untrustworthy systems.

7.1.4 Implications for Practice

The findings presented in this dissertation offer practical implications for the design, deployment, and regulation of AI-based DSSs, particularly in high-risk domains. These implications are especially relevant for practitioners and policymakers interested in responsible human-AI collaboration.

As human users struggle to assess DSS reliability in complex or unfamiliar tasks (even when the system is technically accurate), AI systems should not be deployed uniformly in domains without considering the expertise of the user and the characteristics of the task. Practical implementation should include adaptive interfaces (such as

explanations) or support tools (onboarding and training phases) that help users understand when to question the system. In particular, in light of the findings presented in this thesis, practitioners should consider incorporating brief, guided experiences with the DSS prior to full deployment, allowing users to observe both the system's strengths and its limitations. Where real-time feedback is not feasible (in long-term prediction tasks such as criminal recidivism prediction), designers may consider, if possible, exploring simulated feedback or retrospective cases to approximate the learning benefits of ground truth exposure.

Findings on automation bias (especially the risks posed by initially high-performing but later degrading systems) highlight the need for transparency about system performance over time. Practical DSS designs should incorporate mechanisms that communicate performance variability. Even minimal interventions, such as unspecific alerts about a change in system accuracy, can help users recalibrate their reliance. This approach can be integrated into the interfaces of the system without compromising usability or overwhelming users with technical details.

Finally, we observed that repeated interaction with high-accuracy DSSs may enhance human performance even in unaided decisions, particularly in unfamiliar domains. This implies that AI systems can serve not only as decision aids, but also as informal training tools. Organizations deploying DSSs should explore designs that encourage active engagement rather than passive acceptance, reinforcing learning and judgment over time.

7.2 Limitations

This section describes some limitations of this dissertation, together with future work that can extend and improve the presented research.

7.2.1 Generalizability in Face Matching

We found some limitations during this study, such as approaching more real-world use cases. In use cases for face recognition technologies, the nature of the domain determines under which thresholds of similarity score the machine's response is considered positive or negative. In cases where, for example, it is desirable to prioritize the reduction of false positives (*e.g.*, face recognition methods for private access controls (Ibrahim and Zin, 2011)), the similarity score is set at a higher value than in other scenarios where it is desirable to prioritize the reduction of false negatives (*e.g.*, face recognition methods for law enforcement (Raposo, 2023)). In our work we used symmetric costs. Facial recognition systems are used in complex ethical domains like immigration and law enforcement, where delegating decisions to machines can lead to anonymity, psychological detachment, and invisibility (Ostermaier and Uhl, 2017; Köbis et al., 2019; Hancock and Guillory, 2015). These factors may inadvertently promote unethical actions. It is therefore urgent and necessary to extend research on hybrid systems and their corresponding interaction patterns into domains more closely linked to real-world application fields.

Also, the error behavior of a system is highly dependent on the training data and the architecture of the chosen model. In this work we have chosen two specific pretrained models (IR-50+ArcFace (Deng et al., 2019) and LightCNN (Wu et al., 2018)), well known in the literature, taking care that the training was based on the same dataset MS-Celeb-1M (Guo et al., 2016) for a fair comparison. The error consistency alluded in RQ1 allows us to propose an error characterization, a differentiating axis between human errors and machine errors, on which the exploration and comparisons developed in this work are then based. It is therefore important to bear in mind that in other different scenarios this condition might not be present.

7.2.2 Working with Non-Expert Participants

Our study concerns only two high-risk applications and studies how non-experts interact with a DSS. This allows us to have access to more data points and determine statistical significance of different settings, but reduces the fidelity of the experiment.

Future research should explore adaptive onboarding strategies that consider individual user needs. We also plan to study more deeply the observation done during this experiment that onboarding interact with users' learning of a task. This points out the necessity of experimenting with settings having both training and onboarding phases, potentially at the same time. In high-stakes contexts where human oversight is vital, onboarding can play a pivotal role in ensuring safe and effective deployment of decision support technologies.

7.3 Future Work

This dissertation offers a foundation for understanding human and machine complementarities in decision-making, particularly within high-risk domains such as facial recognition and criminal recidivism prediction. However, several avenues remain open for further investigation. We summarize some of them next.

While the first part of this dissertation empirically establishes the value of human-machine complementarity, further work toward formalizing this into a theoretical model would support field generalization. Such a model would describe under what conditions complementarities are maximized and what cognitive and computational factors drive optimal delegation.

A key limitation of this first part lies in the use of only two pretrained facial recognition models. Future research should replicate

and extend these findings using a broader range of systems, including asymmetric cost scenarios that reflect the priorities of specific domains (e.g., prioritizing false positive reduction in access control or false negative reduction in law enforcement). In parallel, there is a pressing need to examine hybrid human-machine systems in actual operational settings such as border control, policing, and security, where ethical concerns are amplified and interaction patterns may differ due to institutional norms, professional training, and regulatory oversight.

Additionally, investigating how different model architectures, training datasets, and domain-specific biases affect error consistency and human-machine complementarities will provide a more generalizable basis for hybrid decision-making frameworks.

Related to the generalization problem, we believe that future work should involve expert users such as forensic analysts, parole officers, or law enforcement agents to examine whether domain expertise alters reliance patterns, error sensitivity, and trust calibration with DSSs. While the use of non-expert participants facilitated controlled experimentation, it limits the applicability of findings to professional decision-making environments.

Our findings underscore the importance of onboarding phases, particularly when ground truth is available. Future work should develop and evaluate adaptive onboarding strategies that respond to user characteristics (e.g., prior knowledge, cognitive load, confidence levels...) and task complexity. Experiments that vary the timing, duration, and content of onboarding, and experiments that integrate both training and onboarding phases concurrently, may reveal more effective ways to promote appropriate trust and precise mental model of DSSs.

Furthermore, extending onboarding research into longitudinal studies, including also artificial and generic tasks, could illuminate the learning effects of human-machine interaction over time, especially

in less familiar domains. Such studies could also explore whether benefits gained from onboarding persist in future sessions and generalize to similar tasks.

Bibliography

Akhmetova, R. and Harris, E. (2021). Politics of technology: the use of artificial intelligence by us and canadian immigration agencies and their impacts on human rights. In *Digital Identity, Virtual Borders and Social Media*, pages 52–72. Edward Elgar Publishing.

Alon-Barkat, S. and Busuioc, M. (2023). Human–ai interactions in public sector decision making:"automation bias" and "selective adherence" to algorithmic advice. *Journal of Public Administration Research and Theory*, 33(1):153–169.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2022). Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications.

Araujo, T., Helberger, N., Kruikemeier, S., and De Vreese, C. H. (2020). In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI & society*, 35(3):611–623.

Arshad, S. Z., Zhou, J., Bridon, C., Chen, F., and Wang, Y. (2015). Investigating user confidence for uncertainty presentation in predictive decision making. In *Proceedings of the annual meeting of the Australian special interest group for computer human interaction*, pages 352–360.

Aust, J. and Pons, D. (2022). Comparative analysis of human op-

erators and advanced technologies in the visual inspection of aero engine blades. *Applied Sciences*, 12(4):2250.

Baeza-Yates, R. and Estévez-Almenzar, M. (2022). The relevance of non-human errors in machine learning. In *First International Workshop on AI Evaluation Beyond Metrics (EBeM)*. CEUR Workshop Proceedings.

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., and Drachsler, H. (2024). Feedback sources in essay writing: peer-generated or ai-generated feedback? *International Journal of Educational Technology in Higher Education*, 21(1):23.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. (2019a). Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., and Horvitz, E. (2019b). Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 2429–2437.

Bashkirova, A. and Krpan, D. (2024). Confirmation bias in ai-assisted decision-making: Ai triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. *Computers in Human Behavior: Artificial Humans*, 2(1):100066.

Bayram, F. and Ahmed, B. S. (2024). Towards trustworthy machine learning in production: An overview of the robustness in mlops approach. *ACM Computing Surveys*.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models

be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Bigman, Y. E. and Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, 181:21–34.

Bogert, E., Schecter, A., and Watson, R. T. (2021). Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, 11(1):8028.

Bonnefon, J.-F., Rahwan, I., and Shariff, A. (2024). The moral psychology of artificial intelligence. *Annual review of psychology*, 75(1):653–675.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE.

Capdevila, M., Framis Ferrer, B., Soler Iglesias, C., Andres-Pueyo, A., Ruiz Sarrión, L., Arrufat Pijuan, A., Díez Lerma, R., Ribas Plano, P., and Blanch Serentill, M. (2023). Repositori del departament de justícia i qualitat ... - gencat.

Carragher, D. J. and Hancock, P. J. (2023). Simulated automated facial recognition systems as decision-aids in forensic face matching tasks. *Journal of Experimental Psychology: General*, 152(5):1286.

Carragher, D. J., Sturman, D., and Hancock, P. J. (2024). Trust in automation and the accuracy of human–algorithm teams performing one-to-one face matching tasks. *Cognitive Research: Principles and Implications*, 9(1):41.

Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825.

Chandrasekaran, A., Prabhu, V., Yadav, D., Chattopadhyay, P., and Parikh, D. (2018). Do explanations make vqa models more predictable to a human? *arXiv preprint arXiv:1810.12366*.

Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Cheng, H.-F., Wang, R., Zhang, Z., O'connell, F., Gray, T., Harper, F. M., and Zhu, H. (2019). Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12.

Chiang, C.-W., Lu, Z., Li, Z., and Yin, M. (2023). Are two heads better than one in ai-assisted decision making? comparing the behavior and performance of groups and individuals in human-ai collaborative recidivism risk assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Chun, A. H. W. and Wai, H. (2007). Using ai for e-government automatic assessment of immigration application forms. In *AAAI*, pages 1684–1691.

Cummings, M. (2017). Automation bias in intelligent time critical decision support systems. In *Decision making in aviation*, pages 289–294. Routledge.

Davidoff, J., Fonteneau, E., and Goldstein, J. (2008). Cultural differences in perception: Observations from a remote culture. *Journal of Cognition and Culture*, 8(3-4):189–209.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699.

Dietvorst, B. J. and Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological science*, 31(10):1302–1314.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, 144(1):114.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, 64(3):1155–1170.

Dikmen, M. and Burns, C. (2022). The effects of domain knowledge on trust in explainable ai and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162:102792.

Do, H. J., Brachman, M., Dugan, C., Pan, Q., Rai, P., Johnson, J. M., and Thawani, R. (2024). Evaluating what others say: The effect of accuracy assessment in shaping mental models of ai systems. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–26.

Dodge, S. and Karam, L. (2017a). Can the early human visual system compete with deep neural networks? In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2798–2804.

Dodge, S. and Karam, L. (2017b). A study and comparison of human and deep learning recognition performance under visual distortions.

In *2017 26th international conference on computer communication and networks (ICCCN)*, pages 1–7. IEEE.

Dominguez-Catena, I., Paternain, D., and Galar, M. (2022). Gender stereotyping impact in facial expression recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 9–22. Springer.

Dong, M. and Bocian, K. (2024). Responsibility gaps and self-interest bias: People attribute moral responsibility to ai for their own but not others' transgressions. *Journal of Experimental Social Psychology*, 111:104584.

Dratsch, T., Chen, X., Rezazade Mehrizi, M., Kloeckner, R., Mähringer-Kunz, A., Püsken, M., Baeßler, B., Sauer, S., Maintz, D., and Pinto dos Santos, D. (2023). Automation bias in mammography: the impact of artificial intelligence bi-rads suggestions on reader performance. *Radiology*, 307(4):e222176.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

Duan, W., Zhou, S., Scalia, M. J., Yin, X., Weng, N., Zhang, R., Freeman, G., McNeese, N., Gorman, J., and Tolston, M. (2024). Understanding the evolvement of trust over time within human-ai teams. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–31.

Dunning, D. (2011). The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*, volume 44, pages 247–296. Elsevier.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human factors*, 44(1):79–94.

Estévez Almenzar, M., Fernández Llorca, D., Gómez, E., and Martinez Plumed, F. (2022). Glossary of human-centric artificial intelligence. *Sevilla: Joint Research Centre (Seville Site)*.

European Commission (2024). EU AI Act.

Faubel, L., Woudsma, T., Methnani, L., Ghezeljhemeidan, A. G., Buelow, F., Schmid, K., van Driel, W. D., Kloepper, B., Theodorou, A., Nosratinia, M., et al. (2023). Towards an mlops architecture for xai in industrial applications. *arXiv preprint arXiv:2309.12756*.

Feliciano, C. (2016). Shades of race: How phenotype and observer characteristics shape racial classification. *American Behavioral Scientist*, 60(4):390–419.

Flachot, A. and Gegenfurtner, K. R. (2018). Processing of chromatic information in a deep convolutional neural network. *Journal of the Optical Society of America A*, 35(4):B334–B346.

Fontes, C., Hohma, E., Corrigan, C. C., and Lütge, C. (2022). Ai-powered public surveillance systems: why we (might) need them and how we want them. *Technology in Society*, 71:102137.

Frank, M. R., Autor, D., Bessen, J. E., Brynjolfsson, E., Cebrian, M., Deming, D. J., Feldman, M., Groh, M., Lobo, J., Moro, E., et al. (2019). Toward understanding the impact of artificial intelligence on labor. *Proceedings of the National Academy of Sciences*, 116(14):6531–6539.

Fysh, M. C. and Bindemann, M. (2018). Human–computer interaction in face matching. *Cognitive science*, 42(5):1714–1732.

Gao, J., Cao, J., Yeo, S., Choo, K. T. W., Zhang, Z., Li, T. J.-J., Zhao, S., and Perrault, S. T. (2023). Impact of human-ai interaction on user trust and reliance in ai-assisted qualitative coding. *arXiv preprint arXiv:2309.13858*.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., and Ghassemi, M. (2021). Do as ai say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1):31.

Geirhos, R., Meding, K., and Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *arXiv preprint arXiv:2006.16736*.

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., and Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, 34:23885–23899.

Gogoll, J. and Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, 74:97–103.

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, 45:105681.

Green, B. and Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 90–99.

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 87–102. Springer.

Hancock, J. T. and Guillory, J. (2015). Deception with technology. *The handbook of the psychology of communication technology*, pages 270–289.

He, J., Piorkowski, D., Muller, M., Brimijoin, K., Houde, S., and Weisz, J. (2023). Rebalancing worker initiative and ai initiative in future work: Four task dimensions. In *Proceedings of the 2nd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, pages 1–16.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hekler, A., Utikal, J. S., Enk, A. H., Hauschild, A., Weichenthal, M., Maron, R. C., Berking, C., Haferkamp, S., Klode, J., Schadendorf, D., et al. (2019). Superior skin cancer classification by the combination of human and artificial intelligence. *European Journal of Cancer*, 120:114–121.

Hemmer, P., Thede, L., Vössing, M., Jakubik, J., and Kühl, N. (2023). Learning to defer with limited expert predictions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(5):6002–6011.

Holsinger, A. M., Lowenkamp, C. T., Latessa, E., Serin, R., Cohen, T. H., Robinson, C. R., Flores, A. W., and VanBenschoten, S. W. (2018). A rejoinder to dressel and farid: New study finds computer algorithm is more accurate than humans at predicting arrest and as good as a group of 20 lay experts. *Fed. Probation*, 82:50.

Horowitz, M. C. and Kahn, L. (2024). Bending the automation bias curve: A study of human and ai-based decision making in national security contexts. *International Studies Quarterly*, 68(2):sqae020.

Hou, Y. T.-Y. and Jung, M. F. (2021). Who is the expert? reconciling algorithm aversion and algorithm appreciation in ai-supported decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Huegli, D., Merks, S., and Schwaninger, A. (2023). Benefits of decision support systems in relation to task difficulty in airport security x-ray screening. *International Journal of Human–Computer Interaction*, 39(19):3830–3845.

Hupont, I. and Fernández, C. (2019). Demogpairs: Quantifying the impact of demographic imbalance in deep face recognition. In *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–7. IEEE.

Hupont, I., Tolan, S., Gunes, H., and Gómez, E. (2022). The landscape of facial processing applications in the context of the european ai act and the development of trustworthy systems. *Scientific Reports*, 12(1):10688.

Ibrahim, R. and Zin, Z. M. (2011). Study of automated face recognition system for office door access control application. In *2011 IEEE 3rd International Conference on Communication Software and Networks*, pages 132–136. IEEE.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. (2019). Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*.

Jeckeln, G., Hahn, C. A., Noyes, E., Cavazos, J. G., and O'Toole, A. J. (2018). Wisdom of the social versus non-social crowd in face identification. *British Journal of Psychology*, 109(4):724–735.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Jones-Jang, S. M. and Park, Y. J. (2023). How do people react to ai failure? automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, 28(1):zmac029.

Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. In Rowe, F., editor, *28th European Conference on Information Systems - Liberty, Equality, and Fraternity in a Digitizing World, ECIS 2020, Marrakech, Morocco, June 15-17, 2020 : Proceedings*, page RP 168, Atlanta, GA. AISeL.

Kawaguchi, K. (2021). When will workers follow an algorithm? a field experiment with a retail business. *Management Science*, 67(3):1670–1695.

Keding, C. and Meissner, P. (2021). Managerial overreliance on ai-augmented decision-making processes: How the use of ai-based advisory systems shapes choice behavior in r&d investment decisions. *Technological Forecasting and Social Change*, 171:120970.

Keswani, V., Lease, M., and Kenthapadi, K. (2021). Towards unbiased and accurate deferral to multiple experts. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 154–165.

Keyes, O. (2018). The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22.

Kim, S. H., Schramm, S., Riedel, E. O., Schmitzer, L., Rosenkranz, E., Kertels, O., Bodden, J., Paprottka, K., Sepp, D., Renz, M., et al. (2025). Automation bias in ai-assisted detection of cerebral aneurysms on time-of-flight mr angiography. *La radiologia medica*, 130(4):555–566.

Kim, S. S., Watkins, E. A., Russakovsky, O., Fong, R., and Monroy-Hernández, A. (2023). Humans, ai, and context: Understanding end-users' trust in a real-world computer vision application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 77–88.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Köbis, N. C., Verschuere, B., Bereby-Meyer, Y., Rand, D., and Shalvi, S. (2019). Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspectives on Psychological Science*, 14(5):778–796.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. (2020). Big transfer (bit): General visual representation learning. In *16th European Conference on Computer Vision, Part V 16*, pages 491–507. Springer.

Korteling, J. E., van de Boer-Visschedijk, G. C., Blankendaal, R. A., Boonekamp, R. C., and Eikelboom, A. R. (2021). Human-versus artificial intelligence. *Frontiers in artificial intelligence*, 4:622364.

Koulu, R. (2020). Proceduralizing control and discretion: Human oversight in artificial intelligence policy. *Maastricht Journal of European and Comparative Law*, 27(6):720–735.

Kulesza, T., Stumpf, S., Burnett, M., and Kwan, I. (2012). Tell me more? the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 1–10.

Kulesza, T., Stumpf, S., Burnett, M., Wong, W.-K., Riche, Y., Moore, T., Oberst, I., Shinsel, A., and McIntosh, K. (2010). Explanatory debugging: Supporting end-user debugging of machine-

learned programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*, pages 41–48. IEEE.

Küper, A., Lodde, G. C., Livingstone, E., Schadendorf, D., and Krämer, N. (2025). Psychological factors influencing appropriate reliance on ai-enabled clinical decision support systems: experimental web-based study among dermatologists. *Journal of Medical Internet Research*, 27:e58660.

Kyriakou, K. and Otterbacher, J. (2023). In humans, we trust: Multidisciplinary perspectives on the requirements for human oversight in algorithmic processes. *Discover Artificial Intelligence*, 3(1):44.

Laakasuo, M., Palomäki, J., and Köbis, N. (2021). Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics*, 13(7):1679–1688.

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., and Tan, C. (2023). Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 1369–1385.

Lai, X. and Rau, P.-L. P. (2021). Has facial recognition technology been misused? a public perception model of facial recognition scenarios. *Computers in Human Behavior*, 124:106894.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Langer, M. and Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, 123:106878.

Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., and Bermúdez i Badia, S. (2022). Towards efficient annotations for a

human-ai collaborative, clinical decision support system: A case study on physical stroke rehabilitation assessment. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pages 4–14.

Lin, Z., Jung, J., Goel, S., and Skeem, J. (2020). The limits of human predictions of recidivism. *Science advances*, 6(7):eaaz0652.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Longoni, C. and Cian, L. (2022). Artificial intelligence in utilitarian vs. hedonic contexts: The "word-of-machine" effect. *Journal of Marketing*, 86(1):91–108.

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., and Ouyang, W. (2023). Seeing is not always believing: Benchmarking human and model perception of ai-generated images. *Advances in neural information processing systems*, 36:25435–25447.

Lyell, D. and Coiera, E. (2017). Automation bias and verification complexity: a systematic review. *Journal of the American Medical Informatics Association*, 24(2):423–431.

Ma, S., Wang, X., Lei, Y., Shi, C., Yin, M., and Ma, X. (2024). "are you really sure?" understanding the effects of human self-confidence calibration in ai-assisted decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Mahmud, H., Islam, A. N., Ahmed, S. I., and Smolander, K. (2022). What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:121390.

Majidi, F., Khomh, F., Li, H., and Nikanjam, A. (2024). An efficient model maintenance approach for mlops. *arXiv preprint arXiv:2412.04657*.

Makino, T., Jastrzebski, S., Oleszkiewicz, W., Chacko, C., Ehrenpreis, R., Samreen, N., Chhor, C., Kim, E., Lee, J., Pysarenko, K., et al. (2022). Differences between human and machine perception in medical diagnosis. *Scientific reports*, 12(1):6877.

Matias, J. N. (2023). Humans and algorithms work together—so study them together. *Nature*, 617(7960):248–251.

McGee, J. P., Parasuraman, R., Mavor, A. S., and Wickens, C. D. (1998). *The future of air traffic control: Human operators and automation*. National Academies Press.

Meissner, C. A. and Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1):3.

Monahan, J. and Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual review of clinical psychology*, 12(1):489–513.

Monroe, S. and Vangsness, L. (2022). The effects of task difficulty and stress on trust in an automated navigation aid. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 66, pages 1080–1084. SAGE Publications Sage CA: Los Angeles, CA.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., and Zafeiriou, S. (2017). Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59.

Mozannar, H., Lang, H., Wei, D., Sattigeri, P., Das, S., and Sontag, D. (2023). Who should predict? exact algorithms for learning to

defer to humans. In *International conference on artificial intelligence and statistics*, pages 10520–10545. PMLR.

Negri, P., Hupont, I., and Gomez, E. (2024). A framework for assessing proportionate intervention with face recognition systems in real-life scenarios. *arXiv preprint arXiv:2402.05731*.

Neufeld, A. (2017). In defense of risk-assessment tools.

Norman, D. A. (2014). Some observations on mental models. In *Mental models*, pages 7–14. Psychology Press.

Northcutt, C. G., Athalye, A., and Mueller, J. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint 2103.14749*.

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E., and Gogate, V. (2021). Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 340–350.

Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E. D., and Gogate, V. (2022). On the importance of user backgrounds and impressions: Lessons learned from interactive ai applications. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–29.

Ostermaier, A. and Uhl, M. (2017). Spot on for liars! how public scrutiny influences ethical behavior. *PloS one*, 12(7):e0181682.

Papenmeier, A., Kern, D., Hienert, D., Kammerer, Y., and Seifert, C. (2022). How accurate does it feel?–human perception of different types of classification mistakes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Park, J. S., Barber, R., Kirlik, A., and Karahalios, K. (2019). A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–15.

Park, Y. J. and Jones-Jang, S. M. (2023). Surveillance, security, and ai as technological acceptance. *AI & society*, 38(6):2667–2678.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. (2012). Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., and O'Toole, A. J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, 8(2):1–11.

Phillips, P. J. and O'Toole, A. J. (2014). Comparison of human and computer performance across face recognition experiments. *Image and Vision Computing*, 32(1):74–85.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176.

Portela, M., Castillo, C., Tolan, S., Karimi-Haghighi, M., and Pueyo, A. A. (2024). A comparative user study of human predictions in algorithm-supported recidivism risk assessment. *Artificial Intelligence and Law*, pages 1–47.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52.

Prahl, A. and Van Swol, L. M. (2021). Out with the humans, in with the machines?: Investigating the behavioral and psychological effects of replacing human advisors with a machine. *Human-Machine Communication*, 2:209–234.

Prentzas, N., Kakas, A., and Pattichis, C. S. (2023). Explainable ai applications in the medical domain: A systematic review. *arXiv preprint arXiv:2308.05411*.

Pu, P., Chen, L., and Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 157–164.

Punzi, C., Pellungrini, R., Setzu, M., Giannotti, F., and Pedreschi, D. (2024). Ai, meet human: Learning paradigms for hybrid decision making systems. *arXiv preprint arXiv:2402.06287*.

Ramon, M. (2021). Super-recognizers–a novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia*, 158:107809.

Ranjan, R., Bansal, A., Zheng, J., Xu, H., Gleason, J., Lu, B., Nanduri, A., Chen, J.-C., Castillo, C. D., and Chellappa, R. (2019). A fast and accurate system for face detection, identification, and verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(2):82–96.

Raposo, V. L. (2023). The use of facial recognition technology by law enforcement in europe: a non-orwellian draft proposal. *European Journal on Criminal Policy and Research*, 29(4):515–533.

Raposo, V. L. (2024). When facial recognition does not 'recognise': erroneous identifications and resulting liabilities. *AI & SOCIETY*, 39(4):1857–1869.

Reich, T., Kaju, A., and Maglio, S. J. (2023). How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, 33(2):285–302.

Renier, L. A., Mast, M. S., and Bekbergenova, A. (2021). To err is human, not algorithmic–robust reactions to erring algorithms. *Computers in Human Behavior*, 124:106879.

Rice, A., Phillips, P. J., Natu, V., An, X., and O'Toole, A. J. (2013). Unaware person recognition from the body when face identification fails. *Psychological Science*, 24(11):2235–2243.

Romeo, G. and Conti, D. (2025). Exploring automation bias in human–ai collaboration: a review and implications for explainable ai. *AI & SOCIETY*, pages 1–20.

Roy, Q., Zhang, F., and Vogel, D. (2019). Automation accuracy is good, but high controllability may be better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Salehi, P., Chiou, E. K., Mancenido, M., Mosallanezhad, A., Cohen, M. C., and Shah, A. (2021). Decision deferral in a human-ai joint face-matching task: Effects on human performance and trust. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 65, pages 638–642. SAGE Publications.

Salimzadeh, S., He, G., and Gadiraju, U. (2023). A missing piece in the puzzle: Considering the role of task complexity in human-ai decision making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, pages 215–227.

Scaria, A. G., Subramanian, V., George, N. K., and Sengupta, N. (2024). Algorithms and recidivism: A multi-disciplinary systematic review. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1292–1305.

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., and Höllerer, T. (2019). I can do better than your ai: expertise and explanations. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 240–251.

Selten, F., Robeer, M., and Grimmelikhuijsen, S. (2023). 'just like i thought': Street-level bureaucrats trust ai recommendations if they confirm their professional judgment. *Public Administration Review*, 83(2):263–278.

Sharan, N. N. and Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. *Heliyon*, 6(8).

Smiley, L. (2022). 'I'm the operator': The aftermath of a self-driving tragedy. *Wired*.

Sundar, S. S., Waddell, T. F., and Jung, E. H. (2016). The hollywood robot syndrome media effects on older adults' attitudes toward robots and adoption intentions. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 343–350. IEEE.

Szymanski, M., Millecamp, M., and Verbert, K. (2021). Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th international conference on intelligent user interfaces*, pages 109–119.

Szymanski, M., Vanden Abeele, V., and Verbert, K. (2025). Disentangling stakeholder role and expertise in user-centered explainable ai. In *Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, pages 32–39.

The New York Times (2020). `https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html`. Accessed: 2025-05-22.

Towler, A., Dunn, J. D., Castro Martínez, S., Moreton, R., Eklöf, F., Ruifrok, A., Kemp, R. I., and White, D. (2023). Diverse types of expertise in facial recognition. *Scientific reports*, 13(1):11396.

Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., et al. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The lancet oncology*, 20(7):938–947.

Tsipras, D., Santurkar, S., Engstrom, L., Ilyas, A., and Madry, A. (2020). From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*, pages 9625–9635. PMLR.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. (2018). Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*.

Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303.

van Berkel, N., Tag, B., Jacobsen, R. M., Russo, D., Purchase, H. C., and Buschek, D. (2024). Impact of interaction technique in interactive data visualisations: A study on lookup, comparison, and relation-seeking tasks. *International Journal of Human-Computer Studies*, 192:103359.

Van Dijck, G. (2022). Predicting recidivism risk meets ai act. *European Journal on Criminal Policy and Research*, 28(3):407–423.

Wang, M., Deng, W., Hu, J., Tao, X., and Huang, Y. (2019). Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 692–702.

Wang, Q., Zhang, P., Xiong, H., and Zhao, J. (2021). Face.evolve: A high-performance face recognition library. *arXiv preprint arXiv:2107.08621*.

Wang, X. and Yin, M. (2021). Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 318–328.

White, D. and Burton, A. M. (2022). Individual differences and the multidimensional nature of face perception. *Nature Reviews Psychology*, 1(5):287–300.

White, D., Phillips, P. J., Hahn, C. A., Hill, M., and O'Toole, A. J. (2015). Perceptual expertise in forensic facial image comparison. *Proceedings of the Royal Society B: Biological Sciences*, 282(1814):20151292.

Wojcieszak, M., Thakur, A., Ferreira Gonçalves, J. F., Casas, A., Menchen-Trevino, E., and Boon, . M. (2021). Can ai enhance people's support for online moderation and their openness to dissimilar political views? *Journal of Computer-Mediated Communication*, 26(4):223–243.

Wright, D. B. and Sladden, B. (2003). An own gender bias and the importance of hair in face recognition. *Acta psychologica*, 114(1):101–114.

Wu, X., He, R., Sun, Z., and Tan, T. (2018). A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11):2884–2896.

Yang, F., Huang, Z., Scholtz, J., and Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 189–201.

Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12.

Young, A. W. and Burton, A. M. (2018). Are we face experts? *Trends in cognitive sciences*, 22(2):100–110.

Yu, K., Berkovsky, S., Conway, D., Taib, R., Zhou, J., and Chen, F. (2016). Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 223–227.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., and Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces*, pages 307–317.

Yu, K., Berkovsky, S., Taib, R., Zhou, J., and Chen, F. (2019). Do i trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th international conference on intelligent user interfaces*, pages 460–468.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., and Li, Y. (2021). A review of artificial intelligence (ai) in education from 2010 to 2020. *Complexity*, 2021(1):8812542.

Zhang, L., Pentina, I., and Fan, Y. (2021). Who do you choose? comparing perceptions of human vs robo-advisor in the context of financial services. *Journal of Services Marketing*, 35(5):634–646.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. (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*, pages 295–305.

Zhang, Z. T., Argın, S. K., Bilen, M. B., Urgun, D., Deniz, S. M., Liu, Y., and Hassib, M. (2024). Measuring the effect of mental workload and explanations on appropriate ai reliance using eeg. *Behaviour & Information Technology*, pages 1–19.

A Glossary of Human-Centric AI

This appendix is based on Estévez Almenzar et al. (2022) *Glossary of Human-Centric Artificial Intelligence*

A.1 Executive Summary

A.1.1 Policy Context

Over the last few years, Artificial Intelligence (AI) has become a very active field of research, which has evolved from a purely technical field to a research domain spanning different disciplines such as cognitive science, economics, or law. In addition, AI has become a very active topic in terms of policy developments, with governments and institutions defining investment strategies, educational programs or ethical guidelines. The European approach for AI focuses on establishing an ecosystem of excellence and trust in AI in Europe, enabling the development and uptake of AI while ensuring people's safety and fundamental rights. While policy developments in AI should be in line with scientific and technical understanding, there are sometimes differences in vocabulary, which sometimes generate misunderstand-

ings between different research communities or scientists and policy makers.

A.1.2 Key Conclusions

Based on existing literature in the intersection between academia, industry and policy, and given the expertise and know-how developed at the European Commission´s Joint Research Centre (JRC), this work presents a compact but comprehensive glossary of terms on AI, with a focus on a human-centric approach, intended to be used as a relevant reference for interdisciplinary and policy-centred discussions on the topic.

A.1.3 Main Outcomes

We have collected and adapted 230 different terms from more than 10 relevant sources including standards, policy documents and legal texts, as well as multiple scientific references. These include concepts related to trustworthy and human-centred AI such as transparency, fairness or accountability.

A.1.4 Quick Guide

The document is structured as follows. The document first includes a summary of the motivation, goals and structure of the glossary. It then provides the core contribution of the report, which is the list of terms, accompanied by one or several definitions linked to references, and complemented with own definitions when no relevant source was found. The glossary is then complemented by a short discussion on findings, limitations and steps for future work on the topic.

B Contributions

B.1 List of Publications

1. [Baeza-Yates and Estévez-Almenzar (2022)] Baeza-Yates, R. and Estévez-Almenzar, M. (2022). *The Relevance of Non-Human Errors in Machine Learning.* This work was accepted and presented at the Workshop on AI Evaluation Beyond Metrics (EBeM 2022), Vienna, Austria.

2. [*accepted*] Estévez-Almenzar, M., Baeza-Yates, R., and Castillo, C. (2025). *A Comparison of Human and Machine Learning Errors in Face Recognition.* This work was accepted and presented at the European Workshop on Trustworthy AI (TRUST-AI 2025), Bologna, Italy.

3. [*accepted*] Estévez-Almenzar, M., Baeza-Yates, R., and Castillo, C. (2025). *Human Response to AI-Supported Decision-Making in Face Matching: The Influence of Task Difficulty and Machine Accuracy.* This work was accepted as a full paper and presented at the 4th International Conference Series on Hybrid Human-Artificial Intelligence (HHAI 2025). Pisa, Italy.

4. [*submitted*] Estévez-Almenzar, M., Baeza-Yates, R., and Castillo, C. (2025). *Brief Onboarding Phase Improves Deci-*

sion Support: Evidence from Two High-Risk Scenarios. This work was submitted to a journal and is under review.

B.2 Data and Code

- **Chapter 4**
  - Repository: `https://github.com/ealmenzar/Comparison-HumanMachine-Errors-FaceRecognition` (GPL-3.0 license)
- **Chapter 5**
  - Repository: `https://github.com/ealmenzar/HumanResponse-DSS-FaceMatching` (GPL-3.0 license)
- **Chapter 6**
  - Repository: `https://github.com/ealmenzar/BriefOnboaring-DSS` (GPL-3.0 license)

B.3 Other contributions

An additional research project was undertaken during this thesis. Results appear in:

A1. [Estévez Almenzar et al. (2022)] Estévez Almenzar, M., Fernández Llorca, D., Gómez, E., Martinez Plumed, F. (2022). *Glossary of human-centric artificial intelligence*.
This work was carried out during my research stay at the Joint Research Centre (JRC) in Ispra (Italy) from October 2021 to January 2022. It was published by the JRC.