UNIL | Université de Lausanne

Unicentre
CH-1015 Lausanne
http://iris.unil.ch

Year : 2025

Advancing Human-Centered AI Interview Training Grounded in Job Seeker Needs and Responsible Risk Practices

Pooja Shikaripur Bheemasena Rao

Shikaripur Bheemasena Rao Pooja, 2025, Advancing Human-Centered AI Interview Training Grounded in Job Seeker Needs and Responsible Risk Practices

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://iris.unil.ch

Droits d'auteur
L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive IRIS sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright
The University of Lausanne expressly draws the attention of users to the fact that all documents published in the IRIS Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.

UNIL | Université de Lausanne

FACULTÉ DES HAUTES ÉTUDES COMMERCIALES

DÉPARTEMENT DES SYSTÈMES D'INFORMATION

ADVANCING HUMAN-CENTERED AI INTERVIEW TRAINING GROUNDED IN JOB SEEKER NEEDS AND RESPONSIBLE RISK PRACTICES

THÈSE DE DOCTORAT

présentée à la

Faculté des Hautes Études Commerciales
de l'Université de Lausanne

pour l'obtention du grade de
Doctorat en systèmes d'information

par

Pooja SHIKARIPUR BHEEMASENA RAO

Directeur de thèse
Prof. Mauro Cherubini

Co-directeur de thèse
Prof. Dinesh Babu Jayagopi, IIIT Bangalore

Jury
Prof. Valérie Chavez-Demoulin, présidente
Prof. Rafael Lalive, expert interne
Prof. Himanshu Verma, expert externe
Dr. Marios Constantinides, expert externe

LAUSANNE
2025

UNIL | Université de Lausanne
Faculté des hautes études commerciales
Internef - 261
CH-1015 Lausanne

IMPRIMATUR

La Faculté des hautes études commerciales de l'Université de Lausanne autorise le dépôt de la thèse de doctorat rédigée par

Pooja SHIKARIPUR BHEEMASENA RAO

intitulée

Advancing Human-Centered AI Interview Training Grounded in Job Seeker Needs and Responsible Risk Practices

sans se prononcer sur les opinions exprimées dans cette thèse.

Lausanne, le 17.12.2025

Professeure Marianne Schmid Mast, Doyenne

Members of the Thesis Committee

Prof. Mauro CHERUBINI
Faculty of Business and Economics
University of Lausanne
Thesis supervisor

Prof. Dinesh Babu JAYAGOPI
Data Science and AI Department
International Institute of Information Technology Bangalore
Thesis co-supervisor

Prof. Rafael LALIVE
Faculty of Business and Economics
University of Lausanne
Internal Expert

Prof. Himanshu VERMA
Faculty of Industrial Design and Engineering
TU Delft
External Expert

Dr. Marios CONSTANTINIDES
Senior Research Scientist
CYENS Centre of Excellence, Cyprus
External Expert

Advancing Human-Centered AI Interview Training Grounded in Job Seeker Needs and Responsible Risk Practices

Pooja Shikaripur Bheemasena Rao
University of Lausanne

2025

Abstract

Rising unemployment, exacerbated by global economic instability and evolving labor markets, has intensified the need for scalable and effective job preparation tools. Simultaneously, the rapid proliferation of Artificial Intelligence (AI), driven by declining deployment costs and increasing complexity, has fundamentally transformed recruitment practices worldwide. Amid these trends, job seekers face a paradox: while AI offers new avenues for support, it also introduces concerns about unfulfilled user needs, as well as opaque and potentially biased mechanisms of evaluation.

This thesis explores how AI is reshaping the landscape of job interview training while critically examining the risks these technologies pose. Through five interlinked studies, the research traces a progression from designing AI-enabled training systems to examining the broader risks accompanying their adoption in high-stakes settings such as hiring.

The first study introduces InterviewApp, an asynchronous video interview (AVI) training system designed using Self-Determination Theory, a psychological framework that emphasizes fulfillment of autonomy, competence, and relatedness as essential for self-determined motivation and learning. With InterviewApp, I examine how structured training and feedback can enhance job seekers' interview motivation and performance. The intervention with job seekers demonstrates the mediating role of autonomy in influencing performance. Features such as mock interview recordings, playback, and expert feedback positively influence performance. Additionally, the study reveals a strong need for feedback that is personalized, actionable, and available repeatedly, which is challenging to meet through conventional human-led training alone.

The second study employs a participatory co-design methodology to envision how AI might fill this gap. This study collaboratively surfaces needs, constraints, and ethical concerns related to AI-assisted interview training systems by engaging two key stakeholder groups: job seekers and AI practitioners. Findings reveal the requirements job seekers value, their concerns, and practitioner insights into feasibility, pitfalls, and responsible development. It uncovers key tensions between job seekers' perceptions and AI practitioners' expertise around around human–AI collaboration and accountability. The results highlight the four socio-psychological risks perceived by job seekers. These insights offer valuable design considerations for AI interview training systems, demonstrating the significance of stakeholder engagement in AI design for effective user-centered solutions.

The third study builds directly on the concerns raised during co-design, particularly around fairness. I examine how AI, particularly Large Language Models (LLMs), evaluate job interview responses across cultural contexts. Using a controlled dataset of video interview transcripts from UK and Indian job seekers, this analysis examines how LLMs rated attributes such as hirability, self-promotion, storytelling, and positive impression. The findings reveal systematic disparities: Indian transcripts consistently receive lower scores, even when anonymized. Linguistic features, such as syntactic complexity and lexical diversity, rather than identity cues like name or gender, are linked to these disparities. This study highlights the latent cultural biases embedded within LLMs and underscores the need for culturally sensitive evaluation systems in employment settings.

To address the broader implications of such biases, the final two studies interrogate the

wider ecosystem of AI model risk communication and reporting. The fourth study introduces the AI Model Risk Catalog, compiled from nearly 460,000 human-written model documentation. Comparing developer-reported risks with real-world harms from the AI Incident Database and academic risk taxonomies, we identified major reporting gaps, specifically around fraud, manipulation, and social engineering, issues highly relevant to employment-related applications.

The fifth study introduces RiskRAG, a Retrieval-Augmented Generation tool designed to address these documentation shortcomings. Co-designed with AI developers, RiskRAG generates contextual, actionable risk reports tailored to specific models and their applications. User evaluations with practitioners show that RiskRAG improves risk awareness and model selection decisions, fostering more deliberative practices, particularly important in sensitive domains like hiring.

Together, these five studies offer a comprehensive examination of AI's role in job interview training, from system design and participatory development to fairness auditing and systematic risk identification and reporting. This thesis has valuable implications for developing effective tools for job seekers, but also contributes to the broader discourse on the responsible development of AI systems. It emphasizes the need to embed equity, transparency, and accountability across the AI lifecycle, particularly in domains that shape access to economic opportunity.

Published Works

At the time of submission, the second, fourth, fifth, and sixth chapters of this thesis have already appeared in peer-reviewed publications. The second chapter is currently under revision for resubmission. The publications are listed below in the same order as the chapters.

1. **Pooja S. B. Rao**, Laetitia Renier, Marc-Olivier Boldi, Marianne Schmid Mast, Dinesh Babu Jayagopi, and Mauro Cherubini. (2024). On the potential of supporting autonomy in online video interview training platforms. International Journal of Human-Computer Studies, 191, 103326. https://doi.org/10.1016/j.ijhcs.2024.103326

2. **Pooja S. B. Rao**, Laxminarayen Nagarajan Venkatesan, Dinesh Babu Jayagopi, and Mauro Cherubini. Involving Job Seekers and AI Practitioners in Co-designing AI-assisted Interview Training Systems (Under Revision).

3. **Pooja S. B. Rao**, Laxminarayen Nagarajan Venkatesan, Mauro Cherubini, and Dinesh Babu Jayagopi. (2025). Invisible Filters: Cultural Bias in Hiring Evaluations Using Large Language Models. Proceedings of the 2025 AAAI/ACM Conference on AI, Ethics, and Society, 8(3), 2164-2176. https://doi.org/10.1609/aies.v8i3.36703

4. **Pooja S. B. Rao**, Sanja Šćepanović, Dinesh Babu Jayagopi, Mauro Cherubini, and Daniele Quercia. (2025). The AI Model Risk Catalog: What Developers and Researchers Miss About Real-World AI Harms. Proceedings of the 2025 AAAI/ACM Conference on AI, Ethics, and Society, 8(3), 2150-2163. https://doi.org/10.1609/aies.v8i3.36702.

5. **Pooja S. B. Rao**, Sanja Šćepanović, Ke Zhou, Edyta Paulina Bogucka, and Daniele Quercia. (2025). RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 1–26. https://doi.org/10.1145/3706598.3713979

Other relevant, parallel publications, not included in the thesis:

- **Pooja S. B. Rao**, Dinesh Babu Jayagopi, and Mauro Cherubini. (2022). Explainability in Automated Training and Feedback Systems. OSF Preprints. ACM CHI Workshop Human-Centered Perspectives in Explainable AI (HCXAI) https://doi.org/10.31219/osf.io/qd87c

- Kavous Salehzadeh Niksirat, Lahari Goswami, **Pooja S. B. Rao**, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanotham, and Mauro Cherubini. (2023). Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 1–23. https://doi.org/10.1145/3544548.3580848 🏆

- Dhruv Awasthi, **Pooja S. B. Rao**, and Dinesh Babu Jayagopi. (2025). ResumeGenAI: Supporting Job Seekers with LLM-Driven Resume Feedback. Proceedings of the 7th ACM Conference on Conversational User Interfaces, 1–9. https://doi.org/10.1145/3719160.3737622

🏆 Best Paper Award

Acknowledgement

This thesis is the result of a long journey through unknown lands and across unfamiliar peaks. I am grateful to all the people I met along the way and who accompanied me through this process.

First and foremost, I would like to extend my deepest gratitude to my supervisor, Mauro Cherubini. His mentorship has been a cornerstone in shaping me as a researcher. From my early days being new to the field of Human–Computer Interaction to the final stretch of working remotely from India, his guidance has been a constant. The freedom he gave me to pursue my interests and follow my own pace has made all the difference. His ability to ask the right questions at the right moments helped shape the direction of my research in ways I could not have anticipated. His unwavering commitment to research integrity and methodological rigour is something I will carry with me throughout my career. I am equally grateful to my co-supervisor, Dinesh Babu Jayagopi, for his kindness and steadfast support. It is hard to believe that he has been a guiding light to me for a decade now, beginning with my Master's journey. His encouragement is what led me to pursue this PhD in the first place. Thank you for giving me the space to chart the course of my research, for helping shape its direction, and for your advice, patience, and guidance throughout.

I am sincerely grateful to my committee for their thoughtful feedback and their commitment to strengthening this work. I extend heartfelt thanks to Rafael Lalive, whose diverse perspectives, engagement, and financial support through the SNSF fund have been invaluable. I hold deep admiration for Himanshu Verma and Marios Constantinides, for both their intellectual contributions and the rigour and curiosity they exemplify. Their insights have greatly enriched this thesis. Marios has also been a colleague whose guidance and candid insights into the HCI community were invaluable during my internship.

I have been fortunate to have a number of additional mentors. Sanja Šćepanović, you have been a truly amazing mentor. The last part of this thesis would not exist without your support and guidance. Daniele Quercia and Ke Zhou, thank you for your advice and support during my internship and beyond. Edyta Bogucka, thank you for your camaraderie and encouragement. Marc-Olivier Boldi, your statistical and analytical expertise has contributed significantly both to this thesis and to my development as a researcher. Laetitia Renier and Marianne Schmid Mast, your insights and discussions in the early days helped build the foundation of this work. I also thank Tawanna Dillahunt for taking the time to read early versions of my work and offering valuable insights.

I would like to thank my PET lab mates—Kavous Salehzadeh Niksirat, Lahari Goswami, James Arńera, and Alessandro Silacci. Kavous, you have been both a mentor and a friend; I have relied on your advice so often that you have been an invisible advisor throughout my PhD. Lahari, I am grateful that I had a friend like you during the hardest moments of this journey; knowing I had your support meant more than I can express. I am profoundly grateful for all the time and space I shared with you both. James, I have always admired your way of thinking and the uniqueness of your perspectives, something I hope to learn from. Alessandro, your thoughtful suggestions, experiences, patience, kindness, and constant willingness to help are things I will carry with me. This journey has been a thousand times more meaningful and fun because of all of you. I also thank my MPL lab mates, who

supported me in many ways. In particular, I am grateful to Laxminarayanen, not only a collaborator on many projects in this thesis but also someone whose open-mindedness and willingness to work across AI, HCI, and Responsible AI provided a helping hand. I would also like to thank Dhruv, Anastasiia, Sixtine, Jinal, Shreeya, and Vibhav, with whom I had the pleasure of mentoring and collaborating.

My deepest gratitude goes to my family, without whom none of this would have been possible. Amma and Appa, you have been my steady pillars of encouragement in everything I pursue, and you have shaped me into who I am today. To my brother Pavan, who has always been an inspiration. Finally, to my husband, Santhosh. Research can be unavoidably stressful, and you have been my anchor through it all, sharing the weight of this PhD in ways I could never have imagined. As you always say, *Your blood and sweat have gone into this PhD*. Thank you for your presence in my life and for being who you are.

Contents

Abstract . . . iii
Published Work . . . v
Acknowledgement . . . vii

1 Introduction **1**
1.1 Research Context and Motivation . . . 1
1.2 Background . . . 4
1.3 Research Gaps . . . 8
1.4 Research Questions . . . 10
1.5 Thesis Contributions . . . 11
1.6 Thesis Outline . . . 12
1.6.1 Reading This Thesis . . . 14
1.6.2 Author Positionality Statement . . . 14

2 InterviewApp: On the Potential of Supporting Autonomy in Online Video Interview Training Platforms **16**
2.1 Introduction . . . 16
2.2 Background and Hypotheses . . . 18
2.2.1 Model of Interviewee Performance . . . 19
2.2.2 Self-Determination Theory . . . 19
2.2.3 Hypotheses . . . 21
2.3 Related Work . . . 24
2.3.1 Interview Training and Feedback . . . 24
2.3.2 Digital Employment Tools . . . 25
2.3.3 Asynchronous Interview Systems . . . 26
2.4 InterviewApp Design . . . 26
2.4.1 InterviewApp Features . . . 26
2.5 Method . . . 31
2.5.1 Participant Recruitment and Characteristics . . . 31
2.5.2 Procedure . . . 33
2.5.3 Measures . . . 35
2.5.4 Data Analysis . . . 36
2.6 Results . . . 37
2.6.1 Effect on Outcome Variables . . . 38
2.6.2 Mediation Effect on Interview Performance (*H6*) . . . 41

2.6.3 Additional Analysis: Influence on Real-World Job Activities 43
2.6.4 InterviewApp Engagement: Descriptive Statistics 43
2.6.5 Qualitative Interview Results 44
2.7 Discussion 48
2.7.1 Autonomy as a Mediator of Performance 48
2.7.2 Lack of Peer Network 50
2.7.3 Influencing Interviewee Performance 51
2.7.4 Limitations 52
2.8 Conclusion 53

3 Involving Job Seekers and AI Practitioners in Co-designing AI-assisted Interview Training Systems 55
3.1 Introduction 55
3.2 Related Work 57
3.2.1 AI in Interviews and Job Training 57
3.2.2 Stakeholder Involvement in Designing Employment Tools 58
3.2.3 Participatory and Co-Design Approaches in AI Systems 59
3.3 Method 59
3.3.1 Focus Group Workshop with AI Practitioners 60
3.3.2 Co-Design Workshop with Job Seekers 62
3.3.3 Online Evaluation with AI Practitioners 64
3.3.4 Data Analysis 65
3.3.5 Ethics 65
3.4 Findings 66
3.4.1 Findings from Phase 1: Avoiding Pitfalls and Responsible AI Development 66
3.4.2 Findings from Phase 2: Desired Functionalities of Training Systems . 69
3.4.3 Findings from Phase 2: Emerging Risks and Potential Harms 74
3.4.4 Findings from Phase 3: Feasibility and Enhancements 77
3.5 Discussion 79
3.5.1 Alignment on Human-Centered AI for Interview Training 79
3.5.2 Navigating Divergent Perspectives between Job Seekers and AI Practitioners 80
3.5.3 Collaborative Synergies between Stakeholders for Effective AI Training Systems 82
3.5.4 Design Considerations for Automated Training Systems 83
3.5.5 Limitations 84
3.6 Conclusion 84

4 Invisible Filters: Cultural Bias in Hiring Evaluations Using Large Language Models 86
4.1 Introduction 86
4.2 Related Work 89
4.2.1 Cultural Bias in LLMs 89

4.2.2 Bias in Hiring with AI . . . 90
4.3 Method . . . 91
4.3.1 Participant Recruitment and Data Collection . . . 91
4.3.2 LLM-Based Hiring Evaluations . . . 92
4.4 Results . . . 96
4.4.1 Cross-Cultural Differences in LLM Evaluations . . . 96
4.4.2 Influence of Linguistic Features . . . 96
4.4.3 Effect of Identity Cues . . . 98
4.5 Discussion . . . 99
4.5.1 Systematic Cultural Disparities and Consequences for Workforce . . . 99
4.5.2 AI Trust and Overreliance in Global Contexts . . . 100
4.5.3 Toward Culturally Aware LLM Development . . . 100
4.5.4 Limitations and Future Work . . . 101
4.6 Adverse Impact Statement . . . 102
4.7 Ethical Considerations Statement . . . 103

5 The AI Model Risk Catalog: What Developers and Researchers Miss About Real-World AI Harms 104
5.1 Introduction . . . 104
5.2 Related Work . . . 106
5.2.1 AI Risk Taxonomies . . . 106
5.2.2 AI Risks Data Sources . . . 106
5.2.3 Research Gap . . . 107
5.3 Methods . . . 108
5.3.1 Methods for Building the AI Model Risk Catalog . . . 108
5.3.2 Methods for Comparing Risk Sources . . . 110
5.4 AI Model Risk Catalog . . . 111
5.5 Comparison of Risks Envisioned by AI Developers and Researchers with Harms Recorded in the News . . . 114
5.6 Discussion . . . 117
5.6.1 Theoretical Implications . . . 117
5.6.2 Practical Implications . . . 118
5.6.3 Limitations . . . 120
5.7 Adverse Impact Statement . . . 120
5.8 Ethical Considerations Statement . . . 120

6 RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting 122
6.1 Introduction . . . 122
6.2 Related Work . . . 125
6.2.1 AI Risk Documentation . . . 125
6.2.2 Tools for Populating AI Risk Documentation . . . 126
6.3 Deriving Design Requirements From Literature and an Iterative Co-design Process . . . 127
6.3.1 Design Requirements From Literature . . . 128

6.3.2 Design Requirements From Co-design . . . 130
6.4 Designing a Risk Reporting Solution (RiskRAG) Based on the Requirements 133
6.4.1 RiskRAG Architecture . . . 134
6.4.2 Meeting the Design Requirements . . . 137
6.5 Evaluating Our Risk Reporting Solution . . . 138
6.5.1 Baseline RiskRAG Evaluation . . . 138
6.5.2 Preliminary User Study . . . 142
6.5.3 Final User Study . . . 145
6.5.4 Ethics . . . 151
6.6 Discussion . . . 152
6.6.1 Data-driven Solution for Risk Documentation for AI developers . . . 153
6.6.2 Raising Awareness and Promoting Responsible AI Use . . . 154
6.6.3 Broader Implications: Model Cards, Public Outreach, and Policy Making 154
6.6.4 Limitations . . . 155
6.7 Conclusion . . . 157

7 Discussion **158**
7.1 Summary of Contributions . . . 158
7.2 Synthesis . . . 161
7.2.1 Motivation and Autonomy Support in Training . . . 162
7.2.2 Contextualizing Fairness Risks to Uses . . . 163
7.2.3 Participatory Risk Infrastructures . . . 163
7.3 Implications . . . 166
7.3.1 Implications for Job Seekers . . . 166
7.3.2 Implications for AI Practitioners and Developers . . . 167
7.3.3 Implications for Governance and Policy . . . 168
7.3.4 Implications for Risk Documentation Practices . . . 170
7.3.5 Design Recommendations for Future Training and Evaluation Systems 171
7.3.6 Extending Design Implications Beyond Hiring Contexts . . . 172
7.4 Limitations . . . 172
7.5 Future Work . . . 174
7.6 Conclusion . . . 175

References **176**

A Appendix for Chapter 2 **213**
A.1 Mock Interview Questions Selection Procedure . . . 213
A.2 Additional Qualitative Interview Findings: Perception of InterviewApp . . . 213

B Appendix for Chapter 5 **215**
B.1 AI Model Risks Catalog . . . 215
B.2 Comparing risks across sources . . . 218
B.3 Prompts . . . 223

C Appendix for Chapter 6 **226**
C.1 List of papers selected from literature to elicit design requirements for risk reporting 226
C.2 Reporting artifacts generated during the co-design process 227
C.3 RiskRAG and CardGen 227
C.4 Qualitative analysis of feedback from Preliminary Study 228
C.5 Generalizability of RiskRAG to unseen and lesser known models 229

Chapter 1

Introduction

Raam, a temporary sales employee, is seeking a full-time role. With his contract ending soon, the pressure to secure stable employment weighs heavily on him. A short counselling session gave him useful insights, but continuing support was too costly. Practicing with peers proved inconsistent, with their limited knowledge of sales interviews and demanding schedules. He recorded a mock video with a camera and reviewed it with online guides, which offered only vague advice. Despite his efforts, he remained unsure of how to improve and increasingly demotivated.

Raam's experience reflects a broad problem: job seekers often lack access to affordable, scalable training with sustained motivation and actionable feedback. These challenges highlight the need for interview training systems that not only expand access but also provide sustained, fair, and trustworthy support, concerns that are central to this thesis.

1.1 Research Context and Motivation

Rising global unemployment and underemployment remain persistent challenges. According to the International Labour Organization [152], over 183 million people were unemployed in 2024, with youth unemployment exceeding 12% worldwide. As labour markets continue to evolve under the influence of automation and Artificial Intelligence (AI) [235], particularly with the emergence of large language models (LLMs), there is an urgent need to reconsider how we support job seekers in navigating increasingly competitive and digitally mediated hiring processes.

Job interviews remain one of the most widely used methods for candidate selection [307]. However, hiring practices have undergone significant shifts, accelerated by the COVID-19 pandemic [233] and broader trends in automation and digitization [109, 235]. This has led to the widespread adoption of online video interviews, also known as asynchronous video interviews (AVIs) [174], by employers. AVIs are now commonly used in early screening stages[1] and are powered by AI techniques such as computer vision, natural language processing, and audio analysis. AI is becoming increasingly embedded in the hiring pipeline: from sourcing candidates, resume screening, scheduling, to video interview evaluations [104, 114].

[1]For example, `www.hirevue.com`, `www.usebraintrust.com`, `www.viasto.de`, last accessed Aug 2025.

These AVIs are often part of larger Recruitment Management Systems (RMS), which are now in use by a significant majority of organizations. A recent survey by Harvard Business School [114] reported that 63% of all employers and 69% of large enterprises (> 1000 employees) use RMS platforms. Usage is particularly high in the U.S. (75%), compared to 58% in the U.K. and 54% in Germany. Furthermore, over 90% of employers reported using RMS to initially filter or rank skilled candidates.

While AVIs offer clear advantages in terms of efficiency and scalability, they also shift the burden of preparation and performance onto job seekers, often without accessible or reliable guidance. This can be particularly challenging for job seekers from under-resourced or marginalized backgrounds, who may lack the support systems, networks, or professional coaching needed to navigate these high-stakes employment evaluations [95]. Further, AVIs are often designed to filter out candidates, optimize hiring efficiency, and minimize recruiter costs. They are not designed to support job seekers, often reinforcing existing gaps in access and preparedness [114].

At the same time, AVIs also present an opportunity: their underlying technologies can be repurposed to support job seekers through automated feedback and scalable training [173]. Acting as self-help tools, AVIs can reduce dependency on trained counselors and provide repeated practice opportunities. Traditional forms of interview coaching remain limited in reach, leaving many job seekers, especially from marginalized or under-resourced backgrounds, without sufficient preparation or confidence [95]. In this context, AI-assisted interview training offers an opportunity to leverage AVIs as scalable and accessible training tools.

> *On a friend's recommendation, Raam tried an AI-assisted interview training tool. He recorded responses to a set of fixed interview questions and, for the first time, felt some support in practicing on his own. That feeling, however, soon gave way to frustration: the tool offered little beyond a numerical score. No actionable feedback, no tailoring to his role. He was left questioning how the system worked, whether it could be trusted, and whether hidden biases might be shaping its evaluations, just as he had heard about in other AI systems.*

Amid these developments, job seekers face a critical paradox. AI-assisted interview training systems introduce promising new avenues for support, yet questions remain: How can these systems be harnessed to benefit job seekers? What are their expectations and needs from such tools? And how fair and trustworthy are the AI evaluations that increasingly influence their employment opportunities?

To begin addressing these questions, it is necessary to widen the perspective and consider the broader ecosystem of stakeholders involved in AI-assisted interview hiring and training. These systems do not operate in isolation; they involve and impact a range of stakeholders, each bringing different goals, responsibilities, and concerns [88]. These include, along with job seekers, recruiters, AI practitioners and developers, designers, and increasingly, ethics officers, policy makers, and auditors. Given the high stakes involved, it is essential to consider the perspectives and responsibilities of all these stakeholders in the design, evaluation, and governance of AI interview training systems.

Among all stakeholders, job seekers hold the highest personal stake. These systems can

significantly impact access to job opportunities, influencing motivation, self-perception, and livelihoods. However, current AVI-based systems are primarily designed to serve the needs of recruiters, while mostly overlooking the experience and expectations of job seekers [114]. Prior work in digital employment [92, 95, 136] highlights the importance of engaging job seekers in the design and evaluation process. Systems created without job seeker input may increase the gap between what technology offers and what job seekers need, in light of automation, bias, and trust. For these training systems to be effective, they must also promote sustained motivation to prepare [146], helping job seekers stay engaged in training and enhance their performance. Creating optimal environments that foster both skill development and *well-being* is essential; without this, training risks being underutilized or even discouraging.

AI practitioners and developers [13, 70, 213], on the other hand, play a critical role in translating abstract hiring goals into concrete technical systems. Involving these technical experts early in the design phase can foster a more responsible approach to AI development [200, 324]. Their expertise can help anticipate and mitigate potential risks, ethical concerns, and unintended consequences, ensuring that the system operates transparently and fairly [88, 327]. Beyond ensuring functional feasibility, developers are also responsible for aligning system behavior with broader ethical principles and user needs. Critically, their responsibilities extend to the underlying AI models themselves and their application. This includes evaluating and documenting model limitations, reporting foreseeable risks, and communicating how these risks may manifest in both standalone and integrated systems. Responsible model development and risk communication are essential for enabling safe and informed deployment in high-risk domains like employment.

As these AI-assisted interview systems gain influence in the real world, regulators and auditors are increasingly tasked to scrutinize their development and use. According to the EU AI Act (Annex III, pt 3 and 4) [73], AI systems utilized in employment-related decisions, including hiring and vocational training, are classified as *high-risk* [126]. This designation mandates businesses and developers to ensure heightened oversight through rigorous documentation, transparency, and accountability measures. Well-documented systems not only support compliance efforts but also provide accessible risk information to a broad range of stakeholders, including policymakers, auditors, and the general public [113, 149]. For job seekers who directly interact with these systems, transparent documentation helps demystify how AI-generated feedback or evaluations are produced and makes them aware of potential risks of the usage. This can enable more informed engagement, foster trust, and reduce the sense of opacity that often accompanies AI-driven decision-making.

This thesis takes a multi-dimensional approach, foregrounding job seeker needs while critically examining how AI systems are built, evaluated, and documented, to ensure that these tools do not simply serve institutional efficiency, but also promote fairness, transparency, and meaningful support.

1.2 Background

Interview Motivation and Self-Determination Theory

According to the model of interviewee performance [146], motivation is central to the link between interview training and interviewee performance. Interviewee performance is defined as how interviewees behave during the interview, including what they say, do, and how. Prior research has shown that interview-specific self-efficacy, anxiety, and motivation influence performance [146]. This directly affects their capacity to persuade their recruiter of their suitability for that job. Performance in interviews is therefore a function of ability and motivation, which determines whether candidates can effectively translate their skills into successful outcomes [55]. The extent to which interviewees are driven to succeed in their interview could be influenced by their motivation. In this thesis, we adopt the lens of Self-Determination Theory (SDT) [83, 270] of motivation to understand the role of interview motivation.

SDT is a widely used and highly regarded psychological framework for understanding human motivation. Unlike simplistic views that treat motivation as a binary state, SDT defines motivation as a continuum: from a lack of intent (amotivation), through varying degrees of *extrinsic motivation*, to fully self-determined, *intrinsic motivation*. SDT posits that individuals possess these intrinsic motivational tendencies, but maintenance and improvement of these tendencies towards an activity occur when the individual is inherently interested in engaging in the activity. When engagement is driven by *external factors* rather than personal interest, motivation shifts away from intrinsic forms and toward extrinsic motivation on the spectrum. For example, job seekers may prepare for an interview because of peer or family pressure, but they may also be motivated by a clear vision of securing employment or financial stability.

Extrinsic motivation is further differentiated into four types of regulation, each reflecting a different degree of internalization. External regulation represents the least autonomous form, where actions are performed to meet external demands or obtain rewards. Introjected regulation involves behavior driven by internal pressures, such as guilt or anxiety. Identified regulation occurs when individuals recognize the personal value of an activity and willingly accept it as important. Integrated regulation represents the most autonomous form of extrinsic motivation, where the behavior is fully aligned with one's values and sense of self. Finally, true self-regulation is on the extreme end of the spectrum, which is associated with intrinsic motivation. These gradations illustrate how motivation can evolve from externally controlled to more self-determined forms. The full range of motivation is further explained in Chapter 2 and illustrated in Figure 2.2, with examples relevant to job seekers in the context of interview training.

Humans have an innate tendency to grow, master challenges, and seek new experiences, and that drives motivated behaviour and overall *well-being*. The quality of motivation, i.e., its position on this spectrum, depends on whether the environment satisfies or undermines the *basic psychological needs* [270, 271]. Fulfillment of these basic needs leads to more self-determined actions, which makes individuals engage in the activity with *volition*. These needs are autonomy, competence, and relatedness. When fulfilled, they foster intrinsic or

autonomous forms of motivation; when thwarted, motivation tends to be externally regulated or diminished altogether.

Autonomy refers to a sense of volition and choice over one's actions. In the context of interview training, this might suggest volitional engagement and agency over their practice (as we will see in Chapter 2 and 3). **Competence** is the feeling of effectiveness, mastery, and success; this could entail actionable [57] and informational [82] feedback in training (as we will see in Chapter 2 and 3). **Relatedness** reflects the sense of connection and belonging with others. This points to the value of support, whether through peer communities, mentors, experts (as we will see in Chapter 2) or even empathetic system design [266] (as we will see in Chapter 3), that helps job seekers feel less isolated in the training process. Together, fulfilling these needs can create training environments where motivation is sustained, learning is effective, and job seekers are better prepared for interviews.

This theoretical foundation informs Chapter 2 in this thesis, which examines how AVI-based training systems can be designed to enhance both interview motivation and, in turn, performance. By grounding the system in SDT, the study seeks to understand not only whether training improves performance but also how motivation mediates the relationship between training and performance.

Participatory and Co-Design Approaches in AI Systems

A growing movement in AI system development advocates for approaches that go beyond traditional user-centered design to more deeply engage stakeholders in shaping the direction and purpose of these systems: a shift often referred to as the *participatory turn in AI design* [84]. The shift from user-centered to stakeholder-centered design recognizes that AI systems operate within complex socio-technical ecosystems where multiple actors interact [112, 180, 230]. As Forlizzi [112] argues, we are no longer designing for one user, but for multiple stakeholders with overlapping and sometimes competing needs.

Delgado et al. [84] analyzed 80 participatory AI projects and interviewed researchers with experience in participatory methods. They found that while the intent to align AI with stakeholder values is common, actual participation remains shallow. Only 10% of projects empowered stakeholders to influence system goals or define model features. Most projects relied on one-off input methods such as interviews, surveys, or workshops, and only a handful [121, 178] adopted a multi-phase engagement to meaningfully involve stakeholders across the lifecycle. In practice, this means that much of the current participatory AI remains consultative rather than *collaborative*: input is collected, but rarely shapes problem framing, model behavior, or system purpose. Participation thus risks becoming symbolic rather than systemic.

This gap is particularly consequential in high-risk domains such as hiring, where a clear *power dichotomy* exists between the developers and deployers of AI systems (who are referred to as AI practitioners) and the people directly affected by the systems (job seekers, in our case) [198]. For instance, AI-assisted hiring and training systems affect the livelihoods of job seekers who depend on system outcomes. Addressing this imbalance requires rethinking participatory approaches that redistribute power and elevate the voices of those impacted by the system.

Recent efforts have begun to shift towards more inclusive co-design practices. WeBuildAI [180] is a participatory framework that enables community members, organizational staff, and technical facilitators to build algorithmic policy. The approach operationalizes multi-party value deliberation and shows improvements in perceived procedural fairness and outcomes. Bogucka et al. [38] engaged AI practitioners and compliance experts in an iterative co-design process to develop an AI Impact Assessment Report Template. This multi-stakeholder co-design led to a template grounded in regulatory frameworks, which was later evaluated positively by additional practitioners and experts. Kawakami et al. [160] worked with a multi-disciplinary group including AI developers, public sector agency leaders, frontline workers, and community advocates to co-design the Situate AI Guidebook. Through an iterative co-design process, they created a toolkit to support early-stage decision-making for public-sector AI projects.

Grounded in these insights, Chapter 3 of this thesis adopts a multi-phase co-design methodology that moves beyond consultation toward *collaboration*. By engaging both job seekers and AI practitioners across phases, the goal is to align values, operational constraints, and ethical expectations with design decisions. The approach uses tangible artifacts, such as system mockups, feedback prototypes, and sketches, as boundary objects to anchor reflection and feedback. Through this structure, co-design becomes not merely a method but a mechanism to cultivate trust, surface equity-related concerns, and reimagine system functions in ways that would not emerge through unilateral design.

Fairness, Bias, and Risks of AI

As AI systems become increasingly embedded in recruitment and training, fairness and accountability have emerged as pressing concerns. A growing body of work documents how algorithmic hiring tools can reproduce or even amplify existing biases in the labor market [68, 80, 336]. For instance, gender and racial stereotypes present in the workforce have been reflected in hiring algorithms: Facebook's job advertisement system was found to show gender discrimination when promoting job ads exposure[2], and Amazon's infamous automated hiring tool had to be withdrawn after it learned to favor male applicants by mimicking past biased hiring patterns.[3] These risks can be amplified in multimodal settings like asynchronous video interviews, where language, audio, and visual features often encode cultural and demographic differences [250].

A persistent issue is the lack of representative public datasets for studying fairness in hiring contexts, which forces researchers to rely on limited or proprietary data [170]. This has contributed to an excessive focus on gender-based fairness, often neglecting attributes like ethnicity, culture, age, or their intersections [16, 19]. Job seekers themselves highlight a broad set of concerns beyond gender and race to include age, disability, appearance, health, and political beliefs [134]. They also emphasize interaction bias, misinterpretation of qualifications, and the power imbalance of opaque, one-sided AI decisions.

Many vendors of algorithmic hiring tools claim to have solutions for *debiasing* their mod-

[2] https://edition.cnn.com/2025/02/28/tech/facebook-gender-discrimination-europe-ruling-asequals-intl, https://incidentdatabase.ai/cite/416/

[3] https://incidentdatabase.ai/cite/37/

els through techniques like blinding, statistical parity, or compliance with legal thresholds like the 4/5 rule [252]. However, such fixes may create the appearance of fairness without addressing deeper biases or opaque decision-making. The 4/5 rule, which requires that the selection rate of a minority group be at least four-fifths (80%) of that of the majority group, is frequently adopted as a legal safeguard just to avoid a rigorous validation process. Compliance may protect vendors from liability, but it does not resolve structural issues such as differential validity, where an assessment predicts outcomes more accurately for some groups than for others. This gap is further compounded by the mismatch between how computer science treats fairness as a spectrum and how the law treats it as binary (discriminatory or not), placing practitioners under pressure to satisfy both perspectives. Thus, recent work emphasizes that genuine fairness in AI hiring cannot rest on narrow metrics alone; it requires combining technical advances with legal safeguards, stakeholder involvement, socio-technical evaluations, audits, and rigorous documentation [170].

High-risk systems such as AI-assisted hiring and training are explicitly designated under the EU AI Act (Annex III, pt. 3 and 4), mandating heightened oversight through rigorous documentation, risk assessment, and accountability measures [27]. Rigorous risk reporting not only supports compliance but also ensures that risk information is accessible to policymakers, auditors, and the public.

Developers and practitioners carry responsibility not only for designing accurate models but also for clearly communicating the risks of those models, both at the level of technical limitations and in their applied contexts. There are different variations of definitions of risk in the literature. We follow the OECD [244], which defines risk as the likelihood that harm or damage will occur. Recent research has highlighted the importance of systematically identifying and reporting AI risks [70, 88, 324], with calls for structured taxonomies, impact statements, and participatory forms of risk assessment [149, 329]. To address this, researchers have also proposed documentation practices such as Model Cards [213], Datasheets for Datasets [119], and Risk Cards [86], which standardize how systems' intended uses, limitations, and potential harms are disclosed. Yet anticipating risks is difficult for developers, who often lack structured methods or ethical training [44, 98].

Complementing documentation are emerging taxonomies and repositories. Weidinger et al. [329] propose a socio-technical taxonomy of AI risks, distinguishing categories like misinformation, malicious use, human–computer interaction failures, and long-term systemic harms. A meta-review aggregated over 40 prior risk classification frameworks into a unified MIT AI Risk Repository [290], cataloging more than 1600 risks of AI systems. The repository also classifies them by causes and domains such as discrimination, privacy, misinformation, and safety. Alongside taxonomies, incident-tracking initiatives such as the AI Incident Database [207], and the OECD AI Incident Monitor [236] document real-world harms, providing case studies that help practitioners anticipate failures. These resources serve both as checklists for developers and as shared references for auditors, policymakers, and the public.

At the governance level, frameworks such as the NIST AI Risk Management Framework [223] promote lifecycle risk management, urging organizations to map, measure, and mitigate risks continuously. International standards (OECD, ISO) and regulations like the EU AI Act further institutionalize these practices. Together, they signal a shift: risk reporting

is becoming a baseline expectation. For practitioners, adopting taxonomies, repositories, and structured documentation is not only about compliance but also about ensuring that AI systems remain fair, auditable, and trustworthy. Yet, research reveals that only 17% of model cards, the de facto standard for reporting on AI models, address issues related to risks or ethics. Moreover, risk sections in these model cards are often criticized for being overly vague and generic [74]. Addressing this challenge, Chapters 5 and 6 aim to understand the current state of risk reporting of AI models and to assist practitioners and developers in creating actionable and comprehensible AI model risk reports.

1.3 Research Gaps

Interview Training and Motivation. Several research works have examined the influence of interview training on performance, from traditional instructor-led methods [22, 288, 297, 307] to newer approaches using AVIs and virtual agents [78, 173, 218, 291]. A noticeable gap in research is a thorough understanding of how interview training affects the motivation of job seekers. Existing platforms are largely designed around practice and feedback, but do not intentionally promote sustained engagement or motivation. Moreover, these systems have rarely been grounded in psychological theories such as Self-Determination Theory (SDT), which provides a robust framework for understanding how autonomy, competence, and relatedness support intrinsic motivation. Consequently, we lack empirical evidence on whether interview training can satisfy these psychological needs, impact motivation, and improve interview performance.

Stakeholder Involvement in AI System Design. Although participatory design methods have been widely adopted in HCI, their application in AI hiring and training systems remains limited. Previous studies on digital employment technologies emphasized the importance of involving job seekers in design [92, 93, 95, 136]; however, this has not been the case with AI-assisted employment tools, which bring in nuanced needs and fairness concerns. Recent surveys of participatory AI show that stakeholder involvement is often shallow, consultative rather than collaborative, and rarely extends to shaping system goals, models, or features [84, 112]. Multi-phase, iterative engagements where stakeholders influence decisions across the design lifecycle are rare.

Further, current approaches rarely bridge the *power dichotomy* between job seekers and AI practitioners: the perspectives of job seekers are often underrepresented, while AI practitioners' assumptions about feasibility and system constraints remain opaque [198]. As a result, system objectives and functionalities are shaped disproportionately by those in positions of technical or organizational power, leaving limited space for those who are impacted by the outcomes. Bringing both groups into the same design process can surface the tensions between technical feasibility and user expectations, redistribute decision-making power, raise fairness concerns, and jointly shape system goals, features, and safeguards. Involving practitioners early in design is vital, not only to surface system capabilities and limitations, but also to foster more responsible development by anticipating risks, ethical concerns, and unintended consequences [200, 324]. At the same time, systematic approaches that com-

bine job seekers' needs and fairness concerns with the technical expertise of AI developers and practitioners remain scarce. As a result, we lack knowledge of both (*i*) the concrete needs and fairness concerns of job seekers in AI-assisted interview training, and (*ii*) expert practitioner inputs to inform feasibility and responsible design choices.

Fairness and Cultural Sensitivity in AI Hiring. Although fairness in algorithmic hiring has been widely studied, existing research is fragmented and skewed. Most empirical work focuses narrowly on gender or race, often within a single country (primarily the U.S.), leaving other critical axes of inequality such as culture, age, disability, and intersectional attributes largely underexplored [134, 170]. At the same time, work in NLP has demonstrated that LLMs encode latent cultural biases that privilege specific linguistic registers, values, and demographic norms [35]. Yet, cross-cultural disparities in AI-mediated hiring evaluations remain underexplored. Specifically, we lack empirical evidence on whether job seekers from non-Western or underrepresented cultural backgrounds may face invisible disadvantages when their responses are evaluated by AI systems trained predominantly on Western data. Without such evaluation, AI-assisted hiring risks perpetuating inequities by implicitly rewarding cultural-linguistic styles aligned with dominant contexts.

Model-specific AI Risk Reporting. Prior work on AI risk assessment has focused on classifying risks into taxonomies [290, 330], curating harms post-incident [207], or developing tools to automatically generate risk statements for practitioners [49, 324]. However, most existing approaches emphasize researcher-envisioned risks or LLM-generated outputs, with limited attention to the perspectives of AI developers themselves. This overlooks how they perceive unique risk profiles of specific models and their application contexts, and how these concerns align or fail to align with documented harms. We lack an understanding of the current state of risk reporting for AI models from the perspective of AI developers. Further, research reveals that only 17% of model documentation has any risk information [184]. While frameworks like the EU AI Act designate hiring systems as *high-risk*, mandating thorough risk documentation, methods for reporting and communicating such risks to diverse stakeholders remain fragmented and inaccessible. We lack structured approaches to understand the requirements for improved risk reporting and tools to provide support for model-specific concerns, and transparent reporting practices that are actionable in real-world deployment.

Taken together, the identified gaps point to two interrelated areas of inquiry that structure this thesis: (*i*) *Human-centered design and evaluation of AI-assisted interview training systems*, and (*ii*) *Responsible identification and reporting of risks in high-risk AI systems.*

First, we lack evidence on how interview training systems, when grounded in psychological theory such as SDT, can effectively support sustained motivation and improve performance. This highlights a core challenge for human-centered design of AI systems. Second, participatory aspirations in AI-assisted employment systems, and AI systems more broadly, remain shallow. There is limited understanding of how to meaningfully engage both affected users and practitioners to surface needs, preferences, and fairness concerns, and to ensure responsible design choices. Third, fairness research in AI hiring has overlooked cross-cultural disparities, leaving open the question of how such systems evaluate candidates from diverse

backgrounds. Second and third gaps span both areas of inquiry: they require human-centered design and evaluation that gives voice to diverse stakeholders, while also highlighting fairness disparities and stakeholder concerns as risks that must be systematically identified and reported. Finally, given the fairness concerns surrounding high-risk AI systems, there remains a need for systematic approaches to analyse the current state of risk reporting and document and communicate risks in forms accessible not only to developers but also to users, policymakers, and the broader public. This underscores the challenge of responsible identification and reporting of AI risks. Together, the two areas of inquiry frame the scope of this thesis: *understanding how AI-assisted interview training systems can be designed and evaluated to support diverse users, while also developing responsible approaches for identifying and reporting risks that such AI systems may pose.*

1.4 Research Questions

Building on the identified gaps and the two overarching areas of inquiry, this thesis investigates four interconnected research questions. The research questions are designed to address the identified gaps systematically. RQ1 aims to explore the motivational foundations of interview training, focusing primarily on human-centered design. Addressing RQ1 provides a fundamental understanding of how interview training systems can be designed to support motivation and how job seekers engage with them, influencing their performance. Building on this foundation, RQ2 extends the inquiry to the role of AI in making such training scalable. RQ2 examines how both job seekers and practitioners can shape AI-assisted training system design, bridging user needs and concerns with feasibility and responsible practices. In doing so, it connects both areas of inquiry: human-centered design with the responsible identification of risks. Building on this, RQ3 turns specifically to fairness concerns in AI-assisted systems. RQ3 investigates fairness across cultural contexts, asking whether AI hiring evaluations treat candidates equitably and framing potential disparities as both issues of human-centered evaluation and risks that require systematic identification. These fairness-related concerns point to a broader issue: the lack of robust methods for surfacing and documenting risks in AI systems. RQ4 addresses this gap by examining the current state of risk reporting for AI models and exploring how it can be systematically improved to better document and communicate risks, supporting developers.

Together, the four questions form a cumulative research agenda that moves from understanding user motivation to participatory and equitable design to systematic risk identification and reporting.

RQ1. In what ways can interview training systems support job seekers' basic psychological needs for autonomy, competence, and relatedness, and thereby foster motivation and improved interview performance?

RQ2. What are the needs and concerns of job seekers regarding AI-assisted interview training, and how can insights from AI practitioners inform the feasibility and responsible design of such systems?

RQ3. To what extent do AI-based hiring evaluations exhibit bias across cultural contexts, and how might such disparities affect candidates' assessments?

RQ4. What is the current state of AI model risk reporting, and how can it be improved to better support AI developers in documenting and communicating risks?

Addressing these questions requires the involvement of multiple stakeholders: those directly impacted by AI systems and those empowered to design and deploy them. This thesis specifically engages job seekers, AI practitioners, and AI developers. For clarity, the terms AI practitioners and AI developers are used interchangeably to refer to technical experts with hands-on experience in developing, fine-tuning, and/or deploying AI models in real-world applications.

1.5 Thesis Contributions

This thesis makes the following contributions. Each contribution is marked with a label indicating its primary type: **Artifact**, **Empirical**, and **Dataset** according to Wobbrock and Kientz [338].

- Design and development of *InterviewApp*, an asynchronous video interview training tool grounded in Self-Determination Theory (SDT). The system provides structured practice opportunities and feedback on key interview attributes to support job seekers' autonomy, competence, and relatedness. **Artifact**
- A four-month study of *InterviewApp* with job seekers, offering empirical evidence that autonomy significantly mediates interview performance, with implications for the design of future training systems. **Empirical**
- Empirical insights into job seekers' design requirements, preferences, and perceived risks of AI-assisted training, as well as AI practitioner perspectives on responsible development and common pitfalls. **Empirical**
- A cross-cultural study of LLM-based hiring evaluations using interview transcripts from India and the UK, revealing systematic scoring disparities influenced by linguistic features and culturally conditioned preferences. **Empirical**
- The AI Model Risk Catalog, a dataset of 2,863 categorized, unique risks tied to specific models, compiled by analyzing 64,116 model cards with risk sections. This catalog offers a novel resource of developer-reported risks. **Dataset**
- Development of *RiskRAG*, a retrieval-augmented system that generates model-specific risk reports using model cards and incident databases. It supports developers in identifying, contextualizing, and communicating actionable risks. **Artifact**
- Empirical validation of *RiskRAG* reports through studies with AI developers, designers, and end users, showing that it improves argumentation, supports caution in decision-making, and is preferred over standard model cards. **Empirical**

Taken together, these contributions advance understanding of how to design and evaluate AI-assisted interview training systems to support diverse user needs, while also offering broader insights into fairness, cultural sensitivity, and risk reporting in high-risk AI applications.

1.6 Thesis Outline

This thesis is structured around five core chapters, each addressing one or more of the research questions introduced earlier. Together, they explore how AI-assisted systems can support job seekers in interview training, identify fairness and motivational challenges, and propose responsible design and reporting practices for AI systems.

Chapter 2: This chapter presents the design and evaluation of InterviewApp, an asynchronous video interview training system. The design of InterviewApp is grounded in Self-Determination Theory (SDT) and aims to support the three basic psychological needs. It describes the need-supporting features of InterviewApp, including video recording, playback, and structured feedback through peers and expert counselors. The chapter also reports on a four-month study with job seekers, examining the effect of interview training and feedback on job seekers' interview motivation and, in turn, their performance. Findings reveal that autonomy plays a critical role in mediating the effect of training on performance. Engagement with the recording and feedback features positively impacted performance. However, findings also note a lack of supportive peer network for job seekers and limited expert availability for training, underlining the need for more scalable options.

Chapter 3: Informed by the results of Chapter 2, this chapter builds on understanding the needs of job seekers for an AI-assisted interview training system that leverages AI to provide a more scalable training and feedback. This chapter details a multi-phase co-design study involving two key stakeholder groups: job seekers and AI practitioners. It elicits design requirements and perceived risks from job seekers while aligning the system with feasibility and responsible development practices through AI practitioners. The chapter also reflects on the methodological value of progressive engagement in AI system design, especially when power asymmetries exist between users and developers.

Chapter 4: Prompted by the fairness concerns raised by job seekers in Chapter 3, particularly regarding the treatment of diverse candidate groups, this chapter analyzes cultural disparities in LLM-based interview evaluations. This chapter investigates whether large language models (LLMs) used in interview evaluation exhibit cross-cultural bias. It compares 200 interview transcripts from job seekers in India and the UK, examining LLM-generated evaluations across key interview attributes. The chapter analyzes the linguistic features correlated with scoring disparities and explores whether identity-based cues, such as names signaling caste or region, influence evaluations. The findings reveal that Indian transcripts receive consistently lower scores, raising important questions about fairness and cultural sensitivity in AI hiring tools.

Chapter 5: Building on the fairness concerns examined in Chapters 3 and 4, this chapter broadens the focus to the challenge of identifying and cataloging risks associated with AI models used in many high-risk applications. It presents the AI Model Risk Catalog, a

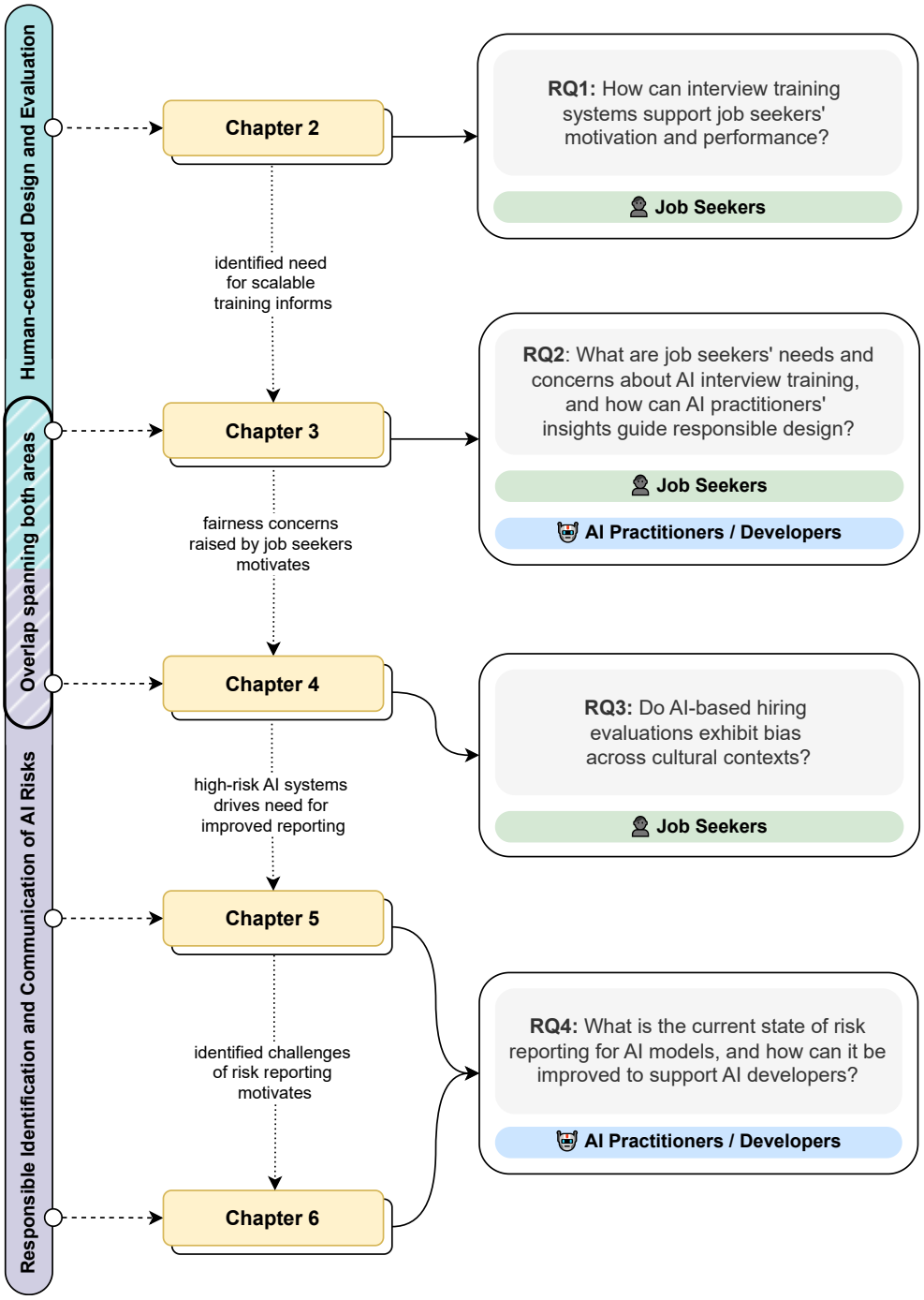


Figure 1.1: Overview of the five thesis chapters, their research questions, and associated stakeholders involved in the study. The chapters progress from job seeker–centered interview training (Chapters 2-3), to examining cultural bias in AI hiring evaluations (Chapter 4), and finally to the identification, communication, and reporting of AI model risks (Chapters 5-6).

dataset of over 2,800 developer-reported risks extracted from publicly available model cards. The chapter details the processes of data collection, categorization, and validation, offering insights into both the types of risks reported and the evolving risk reporting practices of developers. It also compares these developer-reported risks with risks envisioned by researchers and scholars and with harms documented in real-world incidents. By surfacing model-specific risk data at scale, this work contributes to a more transparent and structured foundation for understanding AI model limitations and supporting responsible development.

Chapter 6: Motivated by the finding that developers report risks in only 14% of model cards analyzed in Chapter 5, this chapter introduces RiskRAG, a retrieval-augmented generation system designed to support the creation of actionable and understandable AI model risk reports. Drawing on a co-design study with AI developers, it identifies five key design requirements for improved risk reporting. The chapter then details the architecture of RiskRAG, explaining how it leverages the Risk Catalog and incident databases to generate improved model-specific risk reports. Finally, it presents a multi-stakeholder evaluation involving AI developers, designers, and media professionals. Results show that RiskRAG enhances report clarity, encourages more cautious decision-making, and is consistently preferred over traditional model cards.

Chapter 7: This concluding chapter synthesizes insights across the five studies, drawing out overarching themes related to motivation, autonomy, and risk contextualization. It presents implications for multiple stakeholders, including developers, job seekers, policymakers, and system designers, and reflects on how these findings inform the future design and governance of AI-based hiring systems. The chapter also discusses the broader limitations of the work and identifies directions for future research.

1.6.1 Reading This Thesis

This thesis is organized to reflect how the research program evolved over time: from designing an interview training system, to participatory co-design, to fairness evaluation, and finally to risk reporting and tool development. While this order preserves the developmental logic of the work, several conceptual connections appear across chapters in ways that may benefit from a non-linear reading. For instance, the design insights and stakeholder perspectives presented in Chapter 3 (co-design) provide additional context for understanding the system design choices introduced earlier in Chapter 2; readers may therefore find it helpful to consult Chapter 3 before, during, or immediately after reading Chapter 2. Similarly, Chapter 5 introduces the AI Model Risk Catalog, which forms a foundational component for the system developed in Chapter 6; readers interested in how RiskRAG is constructed may benefit from reading these two chapters together. The outline above highlights these relationships to guide readers through the thesis as both a chronological narrative and an interconnected body of work.

1.6.2 Author Positionality Statement

As a Global South researcher from India conducting a PhD in a Western academic context, my background shapes how I understand AI's role in job interview training and hiring. My

experiences navigating cross-cultural communication and labour markets make me attentive to issues of fairness, representation, and power, particularly when evaluating AI systems developed within predominantly Western paradigms. My engineering and HCI training, along with experience in AI, influence my inclination toward building practical systems and emphasising user-centred design, which shaped the framing and interpretation of Chapters 2 and 3. These experiences also informed my sensitivity to cultural disparities highlighted in Chapter 4, and my focus on responsible and accountable AI practices in Chapters 5 and 6. By acknowledging these influences, I aim to make explicit how my positionality informs the methodological choices and interpretations across this thesis.

Chapter 2

InterviewApp: On the Potential of Supporting Autonomy in Online Video Interview Training Platforms

Abstract. Rising unemployment has led to many discouraged job seekers. While the impact of job seekers' motivation on interview performance is acknowledged in previous research, little attention has been given to understanding the effect of training on interview motivation *and* performance. We present `InterviewApp`, an online interview training tool aiming to support interview motivation through autonomy, relatedness, and competence needs derived from Self-Determination Theory and, in turn, performance. Through a four-month study (N=135), we assess its effectiveness in supporting job seekers' interview motivation and performance. Our results demonstrate the role of autonomy in mediating the effect of training on performance. We found that the intervention significantly affected the job seekers' perceived autonomy. Furthermore, engagement with the recording and feedback features of the tool positively impacted performance. Overall, job seekers found InterviewApp helpful for online interview training and valued the provided expert feedback. These findings have implications for the design of online interview training tools and for behaviour change interventions to support employment.

2.1 Introduction

Rising unemployment is one of the critical issues plaguing many countries. The economic downturn has deteriorated labour market conditions for young people with less than tertiary education [231]. This issue has worsened since the COVID-19 outbreak. Long-term unemployment has increased, and its aftereffect has not been the same for all. [233]. While OECD nations experienced job growth after the pandemic, labour force participation, employment rates, and employment in low-paying and low-skilled jobs are below pre-crisis levels in certain countries [234]. In these situations, it is critical to provide support and training to job seekers during the unemployment spells.

Furthermore, businesses are evolving in ways that quicken the pace of established trends like automation and digitization [233]. Job interviews are the most frequent selection method [307], and hiring new employees through digital platforms is becoming increasingly popular [109]. Although online interviews were a necessity during the pandemic [156], they have now become a convenient and efficient choice for both employers and candidates alike [102, 267]. Conducting job interviews online poses a variety of obstacles to applicants' job prospects [157], warranting the necessity for online interview training.

A number of research works in the literature discuss the topic of interview skills training and its validity in influencing interviewee performance (e.g., [142, 218, 297, 307]). These works often rely on traditional instructor-led training methods. These methods are labour- and time-intensive and dependent on the availability of qualified practitioners in the field. This solution cannot be scaled up to accommodate more training sessions due to challenges associated with scheduling and counsellor availability. Human-Computer Interaction and technology-related research can help tackle these challenges and play a key role in supporting job seekers.

Newer training methods involving technologies like gaming and virtual agents [77, 173, 302, 348], offer chances to boost efficiency and cut expenses. Asynchronous video interviews (AVI), also called online video interviews, [174], provide a rare opportunity to reduce the dependency, to an extent, on the availability of trained practitioners for interview training. Currently, AVIs are used for pre-selection or screening during recruitment.[4] Leveraging them as a self-help tool can provide job seekers with the needed support and reduce the burden on the trainers. As there is a time gap between the interview recording and its evaluation, AVIs can lessen scheduling and financial burdens. The anytime, anywhere feature of AVIs can make training accessible and inclusive.

Prior research has indicated that the mindset of the candidates coming into the interviews can have a large effect on the outcomes of the selection process. Interview-specific self-efficacy, anxiety, and motivation have the potential to influence interviewee performance [146]. Further, previous literature mostly analyzes the direct effects of training on interviewee performance. Interviewee motivation is one of the factors directly linked to an interviewee's capacity to navigate the emotional and interpersonal presentation elements of the interview [146]. Only a few studies have investigated interviewee motivation as related to interviewee performance [173, 205]. Given the belief that performance is a function of ability and motivation [55], it is critical to understand the influence of interview training on interviewee motivation and, in turn, on performance. Additionally, it is crucial to observe participants' changes in interviewee motivation and performance over time as a result of training [173], which is missing from earlier studies.

The main focus of the paper is to assess how effectively interview training and feedback can support job seekers' interview motivation and, in turn, their interview performance. To understand this, we design and evaluate an asynchronous video interview training tool named `InterviewApp`. We rely on the Self-Determination Theory (SDT) [271] to conceptualize interview motivation. SDT posits that satisfying the basic psychological needs of *autonomy* (feeling in control of one's behaviour), *competence* (feeling a sense of mastery), and *relatedness* (feeling a sense of belonging) leads to more self-determined motivation. Consequently,

[4]For example, `www.hirevue.com`, `www.modernhire.com`, `www.viasto.de`, Last accessed August 2025.

the design of the InterviewApp aligns with these principles and is aimed at supporting the three basic needs. InterviewApp includes features such as record, playback, and feedback from an expert or peer. Using ratings and recommendations that can be acted upon, InterviewApp offers structured and thorough feedback on attributes that are thought to be crucial in a job interview. The questions that drive this research and guide the evaluation of InterviewApp are as follows:

RQ1. Does the feedback from InterviewApp support job seekers in increasing their perceived competence at interviews?

RQ2. Does the feedback help improve their perception of job interview relatedness?

RQ3. Does higher engagement on InterviewApp with more practice interviews help job seekers in achieving higher levels of autonomy?

RQ4. Does the feedback from InterviewApp help improve job seekers' interview performance?

In a four-month long study involving 135 participants with a mixed-methods design setup, we compared a control group with experimental groups having access to different features of InterviewApp. We found that access to expert feedback notably increased autonomy compared to playback-only access. The self-determined act of requesting feedback positively influenced autonomy. Furthermore, we identified that autonomy mediates the effect of training on interview performance. Job interview training exhibited a discernible influence on interview performance, with the degree of influence being contingent on the level of autonomy perceived by the job seeker. Contrary to our expectations, findings indicate this relation between autonomy and performance to be *negative* (i.e., the higher the autonomy, the lower the performance). Additionally, engagement with features like record, playback, and feedback positively influenced job seekers' performance.

In this work, we make the following contributions:

- Design of an online video interview training tool informed by a theory of human motivation, providing feedback on various attributes considered essential in an interview to support job seekers.
- Empirical results of a large-scale four-month study. Our results demonstrate the role of interviewees' autonomy in influencing their performance.
- Implications for the design of asynchronous video interview training tools to better support job seekers in their job search.

Our work calls for more behaviour change research into digital employment tools to support job seekers during their unemployment spells.

2.2 Background and Hypotheses

In this section, we provide a brief primer on the theoretical framework on which we build our hypotheses and design and implement the InterviewApp study. Following this, we discuss the development of the six hypotheses.

2.2.1 Model of Interviewee Performance

Huffcutt et al. [146] developed a conceptual model of interviewee performance. Interviewee performance, in this model, is positioned as the central variable between interviewee attributes (e.g., education and experience) and interviewer ratings. Accordingly, interviewee performance can be defined as how interviewees act during the interview and what they say and do. The content of interviewee responses, how they convey the content (e.g., articulation and pitch), and nonverbal behaviour (e.g., posture) are all included in interviewee performance. According to the model, the interviewee state—interview specific motivation, self-efficacy, and anxiety—is one of the factors that affect interviewee performance.

Further, Huffcutt et al. [146] suggests that interview training can directly alter the interviewee state and interviewee performance. Concerning the effect of training on interviewee state, Maurer et al. [205] found that voluntary prior preparation and coaching for interview skills were positively related to interview motivation. Huffcutt et al. [146] define interview motivation as the degree to which interviewees are motivated to do well in interviews. Concerning the effect of training on performance, several studies have demonstrated this [77, 142, 297] (we will come back to this in 2.3.1). Thus, training might influence state (i.e., interview motivation, self-efficacy, and anxiety) and might further influence interviewee performance. Finally, this cause-and-effect chain influences interviewer ratings.

Interviewer ratings are defined as the operational measure of interviewee performance, with the interviewer evaluating the performance and forming the final component of the model as shown in Fig. 2.1. In the context of this paper, interviewer ratings are the subsequent real-world evaluation of interviewee performance in an interview that the job seeker has taken outside of the current study.

The main focus of this paper is derived from the cause-and-effect chain between interview training, interviewee state, and interviewee performance (see Fig. 2.1). We evaluate how interview training influences interview motivation, self-efficacy, and anxiety, and hence interviewee performance.

2.2.2 Self-Determination Theory

Motivation appears central to the link between interview training and interviewee performance. To better understand its role, we rely on the Self-Determination Theory (SDT) [83]. SDT is a broad theoretical framework used to study human motivation and development. SDT assumes that humans have an innate tendency to grow and master challenges, and seek out new experiences, which drives motivated behaviour. SDT views motivation as a continuous spectrum ranging from passive compliance to active personal commitment. The different forms of motivation lie on a continuum of self-determination: complete lack of intent to act (amotivation — non-regulation) at one end and autonomous choice of doing an activity for inherent satisfaction (*intrinsic motivation* — intrinsic regulation) at the other.

Extrinsic motivation spans the spectrum between amotivation and intrinsic motivation, ranging in how autonomously the behaviour is regulated (i.e., relative autonomy). For example, job seekers may be extrinsically motivated to prepare for an interview owing to peer pressure, but they may also be motivated because they envision a valuable consequence (i.e., getting a job or financial gains). Extrinsically motivated behaviours can be classified into four

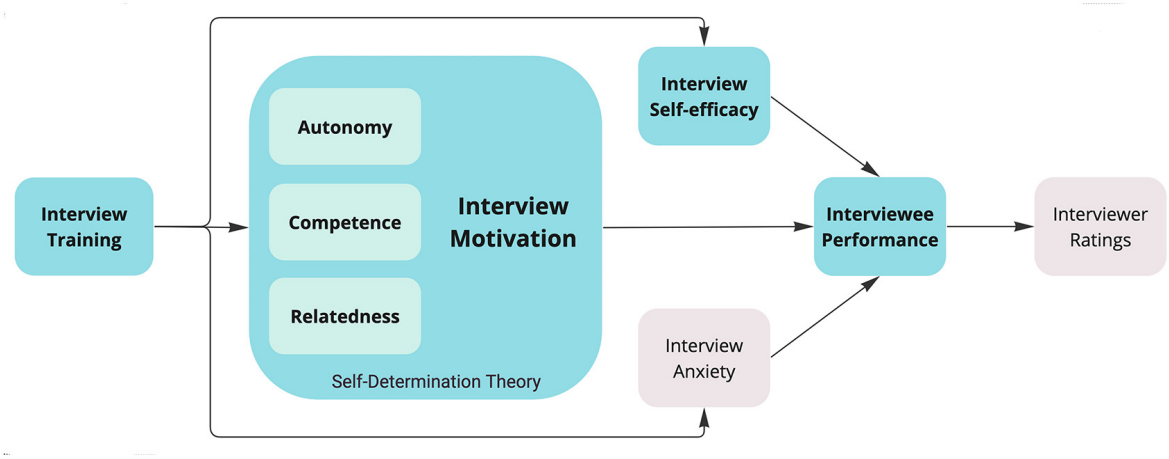


Figure 2.1: Effect of interview training on interview motivation and performance. The constructs in bold (autonomy, competence, relatedness, self-efficacy, and performance) are the main focus of this study and are included in the hypotheses. The constructs of anxiety and interviewer ratings are a part of the exploratory analyses of this study (see Sec. 2.6.3). The effect of interview motivation on training is not depicted, as it is beyond the scope of this study (see Sec. 2.5.2.4).

	Amotivation	Extrinsic Motivation				Intrinsic Motivation
Types of regulation	**Non regulation**	**External regulation**	**Introjected regulation**	**Identified regulation**	**Integrated regulation**	**Intrinsic regulation**
Description	Lacking intent to act	Performed to satisfy external demand or reward	Performed to avoid guilt or please others	Performed as personally accepted action	Performed in congruence with one's values	Inherent satisfaction
Examples in the context of interview training	Unmotivated to practise for uninteresting jobs	Adapting to online format of interviews and train for remote positions	Prepare for an interview owing to soceital pressure	Train for interviews to get a job as it is an important part of life	Train for personal growth and showcasing one's unique abilities to secure meaningful job opportunities	Train to acquire transferable skills for enhanced job performance and long-term career growth, beyond job acquisition

Figure 2.2: The continuum of human motivation ranging from amotivation to intrinsic motivation (adapted from Ryan and Deci [270]). The lower panel gives examples corresponding to the context of this study for each type of motivation.

types : those undertaken to meet external demand (external regulation), those performed to avoid guilt or anxiety (introjected regulation), those recognized or owned as personally meaningful (identified regulation), and those fully integrated into the self (integrated regulation). Finally, true self-regulation is on the extreme end of the spectrum, which is associated with intrinsic motivation [270]. Fig. 2.2 illustrates this spectrum of motivational types with examples in the context of job interview training.

Basic Psychological Need Theory [270, 271], one of SDT's mini-theories, characterises basic psychological needs (BPNs) —*autonomy, competence* and *relatedness* —as something that needs to be satisfied for an individual's integrity, development, and well-being.

- *Autonomy* is "a sense of initiative and ownership in one's actions. It is supported by experiences of interest and value and undermined by experiences of being externally controlled, whether by rewards or punishments." [272, p. 1]
- *Competence* is "the feeling of mastery, a sense that one can succeed and grow. The need

for competence is best satisfied within well-structured environments that afford optimal challenges, positive feedback, and opportunities for growth." [272, p. 1]
- *Relatedness* is "a sense of belonging and connection. It is facilitated by the conveyance of respect and caring." [272, p. 1]

The fulfilment of the three basic psychological needs described above has been found to be conducive to self-determined motivation [270]. In turn, self-determined motivation has been linked to higher effort, persistence, and performance levels towards a target activity (i.e., preparing for job interviews) [270, 314]. Job seekers driven by self-determined motivation tend to participate in more training than job seekers driven by extrinsic factors (e.g., peer pressure) [315]. Going through an unemployment spell with self-determined motivation likely contributes to securing high-quality and longer-lasting employment [315]. Consequently, a *circular relationship* emerges between adherence to training and motivation to perform well in job interviews: job seekers who invest more effort in their interview training end up better fulfilling their basic psychological needs, which provides them with self-determined motivation towards performing well in interviews. This usually involves training more for the upcoming job interviews, which again fulfils their basic psychological needs, and so forth. We probe the effect of training on motivation and performance, while the reciprocal effect of motivation on training is beyond the scope of the current study design.

As noted in Ryan and Deci [270], humans possess intrinsic motivational tendencies. However, maintenance and enhancement of this inherent propensity requires supportive conditions. Our study aims to nurture and sustain this self-determined motivation to perform well in interviews by creating supportive conditions during training for forthcoming job interviews. Fostering autonomy, competence, and relatedness during the training, the InterviewApp tool aims to create an environment that can help job seekers experience greater fulfilment of their basic psychological needs and nurture their self-determined motivation to perform well in interviews, as well as their performance during the actual interview. However, it is important to note that the outcome of a job interview is influenced by many factors, most of which are not in the control of the job applicant and lie beyond the purview of this study (e.g., the performance of the other job seekers, fit for the job, as assessed by the recruiter, interviewer behaviour).

Fig. 2.1 is a visual depiction of the conceptual framework integrating the SDT with the previously presented model of interviewee performance. We adapt this framework to develop our hypotheses and guide the present study.

2.2.3 Hypotheses

Deci and Ryan assert that positive feedback and praise boost intrinsic motivation (as one's perception of competence grows) [83]. They note that all feedback has two functional properties. One is the controlling aspect—the target activity is done to obtain an external reward or to conform to an external restriction. The other is the informational aspect—enables opportunities and offers knowledge to help with the activity. Informational feedback is key to positive competence and increased intrinsic motivation. If the feedback provides individuals with helpful information that will enable them to engage with the environment, influencing the activity, it can be considered informational [82, 83]. The information in the feedback

should be actionable to be incorporated by the individual. Actionable feedback is characterized by its ability to facilitate both learning and tangible outcomes [57]. Competence is facilitated by feedback, both positive and negative, containing information that might help with problem-solving. Negative feedback can provide hints toward future competence rather than suggesting ineptitude [83]. Through experts and peers, InterviewApp aims to provide informational feedback that can help job seekers perform effectively in a job interview (see Sec. 2.4.1.4 for details). Thus, we pose our first hypothesis:

H1 : *Job seekers receiving feedback will have increased competence compared to job seekers receiving no feedback.*

Furthermore, we believe that the feedback from an expert trained to evaluate interviews will be high in the informational aspect. Hence, the following sub-hypotheses.

H1a : *Job seekers receiving feedback from experts will have a higher improvement in competence compared to job seekers receiving feedback from peers.*

H1b : *Job seekers receiving feedback from peers will have a higher improvement in competence compared to job seekers receiving no feedback.*

Recognition, appreciation, or a well-timed remark from someone can make a person feel significant [271]. It can make a person feel more relevant and connected to the other person. Three elements are crucial to support relatedness: (i) *acknowledgement*–being acknowledged by another person, having one's full attention, and being listened to; (ii) *support*–beyond acknowledgement, desire to be supported and respected; and (iii) *impact*–seeing an impact on others [266]. This experience of relatedness can be encouraged by establishing connections to an individual or a network of people via a technology-assisted medium [265, 266]. With the potential to link job seekers to qualified experts and peers through feedback, we intend to foster relatedness via the elements of *acknowledgement* and *support*. Feedback on interview performance from an expert or a peer can offer recognition and recommendations, increasing job seekers' perception of acknowledgement and supporting their abilities. Therefore, we pose our second hypothesis:

H2 : *Job seekers receiving feedback will have increased relatedness compared to job seekers receiving no feedback.*

Additionally, a job seeker may feel more valued and appreciated when a qualified expert provides feedback and recognition on their interview performance. Furthermore, compared to feedback from a friend, evaluation from a stranger will be seen as more objective, with increased perception of support [94, 96], boosting relatedness. Therefore, we hypothesize further:

H2a : *Job seekers receiving feedback from experts will have a higher improvement in relatedness compared to job seekers receiving feedback from peers.*

H2b : *Job seekers receiving feedback from peers will have a higher improvement in relatedness compared to job seekers receiving no feedback.*

The characteristics of choice and opportunity are associated with autonomy satisfaction [272]. Autonomy is supported by designs that provide individuals with the choice to pursue the problems and activities they wish to take on, as well as the tactics and solutions they will employ to do so. Another path to autonomy satisfaction exists even in settings with fewer choices. Having a sense of purpose and acting in accordance with one's principles and interests can lead to autonomy. Employing *compelling rationales* to stimulate an internal

desire to follow the task at hand, even in settings with limited choice, leads to autonomy and is known as *volitional engagement* [266]. The long-range goal of succeeding in a job interview provides a sense of meaning for job seekers, promoting volitional engagement. We intended to design InterviewApp with a combination of meaningful choices and volitional engagement (see Sec. 2.4 for details). Participants were free to choose when and how to use InterviewApp during the study. Hence, the following hypothesis:

H3 : *Job seekers with higher engagement with InterviewApp during the experiment will report higher autonomy levels.*

Huffcutt et al. [146] posit that interview training can have a positive impact on interview self-efficacy. Perceived self-efficacy is defined as "judgments of how well one can execute courses of action required to deal with prospective situations" [20] and in this context, judgment of one's ability to perform well in an interview. When combined with self-reflection, assessment can boost self-efficacy in skill acquisition [339]. Self-reflection refers to a form of self-focus directed at oneself that is driven by a sense of curiosity or a desire to acquire knowledge about oneself [301, 310]. Providing opportunities for self-reflection and self-evaluation can enhance job seekers' understanding of their current performance and identify areas for improvement, thereby increasing their belief in their interview abilities. Previous studies have demonstrated that self-reflection and feedback facilitated by various digital employment tools can enhance self-efficacy [69, 91, 94]. InterviewApp provides an option to playback interview videos to self-review (see Sec. 2.4.1.3) and get feedback, aiming to facilitate self-reflection among job seekers. Consequently, we pose our fourth hypothesis:

H4a : *Job seekers who practise interviews with self-review will increase their interview self-efficacy compared to job seekers with no self-review.*

H4b : *Job seekers receiving feedback will have increased self-efficacy compared to job seekers receiving no feedback.*

As discussed in the model of interviewee performance (Sec. 2.2.1), training and feedback can influence interviewee performance. InterviewApp provides evaluation and feedback on seven measures that are key to performing well in a job interview [147]. Thus, we pose our fifth hypothesis:

H5 : *Job seekers receiving feedback will report higher levels of improvement in performance measures than job seekers receiving no feedback.*

Further, job seekers might consider the feedback from a qualified expert more seriously than a peer and work towards incorporating it, having an impact on their interview performance. Hence, we hypothesize further:

H5a : *Job seekers receiving feedback from experts will have a higher improvement in performance measures compared to job seekers receiving feedback from peers.*

H5b : *Job seekers receiving feedback from peers will have a higher improvement in performance measures compared to job seekers receiving no feedback.*

Following the combined framework in Fig. 2.1, we seek to understand the mediating effects of autonomy, competence and relatedness on performance. According to the cause-and-effect chain between interview training, interviewee state, and interviewee performance in the model of interviewee performance, interview motivation acts as a mediator between interview training and performance. Huffcutt et al. [146] emphasized the need for further investigation into the mediating effects of interview motivation on performance. Thus, we

intend to understand if the three BPNs serve as mediators of interview performance and pose our last hypothesis[5] without presupposing a specific direction of influence:

H6 : *The three basic psychological needs of autonomy, competence, and relatedness will mediate the effect of interview training on interviewee performance.*

2.3 Related Work

This section will begin by reviewing some of the literature on interview preparation and feedback, followed by the literature related to different digital employment tools. Finally, we will discuss a few works related to AVIs.

2.3.1 Interview Training and Feedback

Various studies investigate the effects of training on improving interview performance. A previous review of literature in job interview training [176] concluded that performance gains across targeted interview skills were observed in most of the training methods reviewed. They contended that—instruct, model, practice, and feedback, with minor variations known as Behavioural Skills Training (BST)—were the most widely used and efficient training techniques. Hollandsworth et al. [142] performed a study with college students to assess the relative effectiveness of an interview skills workshop using BST compared to a lecture-discussion workshop. They conclude that the most promising method for training job interview skills is a model that combines the most effective elements of the behavioural and discussion groups. Stocco et al. [297] conducted a single-subject analysis study on five college students to evaluate BST. They found that all students acquired the targeted skills and were satisfied with the training outcomes.

Novel methods of technology provide opportunities for training with more flexibility and reduce scheduling conflicts and expenses. A few studies explore the possibility of using these methods for job interview training. Langer et al. [173] introduced and evaluated virtual interview training in comparison with the classical approaches of training, including sample questions and video presentations. They found that virtual training improved interview performance and participants' hireability. Muralidhar et al. [218] designed and implemented a behavioural training procedure for hospitality students to train for social skills in a job interview scenario, leading to improved performance.

There have also been attempts to add automated elements to training and feedback systems. `MACH` [143], `TARDIS` [14] are coaching-based conversational agents. They focus on skill assessment and provide non-verbal behaviour analysis as feedback to interviewees to improve the conversational and social skills of the participants. MACH is a virtual social skill coach in the context of job interview training. `ASST` [303] focuses on social skills training as its interview objective. `Parley` [282] is an interactive system intended to train participants in social situations like job interviews.

[5] *H6* was not a part of our preregistration. However, we added it as a part of our data analysis protocol at the end of the data collection and before the actual analysis was performed. All supplementary materials of the paper are available on OSF at https://doi.org/10.17605/osf.io/b9qfj. Deviations from the preregistration are explained in Sup. 12 .

However, the majority of these studies analyse the direct effects of training on performance rather than its influence on the behaviour and mindset of job seekers. A finer-grained analysis to uncover the specific components through which interview training and feedback impact motivation and, consequently, interview performance is missing in the previous studies. We believe the SDT can provide a strong theoretical framework to gain insights into these components and help understand the long-term effects on interviewee motivation and performance. Furthermore, while many of these studies assess the impact of training in simulated environments or with student participants, our study focuses on evaluating the effectiveness of InterviewApp with actual job seekers in real-world settings.

2.3.2 Digital Employment Tools

In order to address the needs of underemployed workers, Dillahunt et al. [95] developed ten concrete design principles. They categorize job seeker needs and challenges into social (needs met by social connections and ties), personal (needs requiring effort from oneself, such as reflection and skills), and societal (needs requiring support from the community or government), and transform them as design concepts for digital unemployment tools. They formulated ten design concepts to address these three needs, for example, `Interview4`, `Review-Me`, `DreamGigs`, etc. In a speed-dating survey of these ten interventions, they found that low-income job seekers choose resources that meet their more urgent and not long-term unemployment needs.

Our work is closely related to the recent study by Dillahunt and Hsiao [94]. They conducted a one-month long study to assess digital unemployment tools'— Review-Me and Interview4 — effectiveness in supporting underrepresented job seekers using the three factors influencing job seeker intentions to search for jobs—job search self-efficacy, subjective norms, and job search attitudes. The study revealed that the tools showed positive trends in job seekers' self-efficacy and subjective norms. Features that integrated aspects of positive feedback and self-reflection were effective in improving job seekers' self-efficacy. In a replication study of [94] on underrepresented job seekers in Switzerland [96], it was discovered that treatments that provided support to self-regulation could help job seekers in constrained environments. The authors found that external regulations like those imposed by job placement institutions impacted job seekers' autonomy. Hence, the authors using the SDT paradigm note that interventions that provide self-regulation training could help support job seekers' autonomy to achieve their goals.

The online video interviewing platform InterviewApp, developed as part of this work, can be seen as an extension of the previously devised Interview4. InterviewApp is intended to serve the personal and social needs of job seekers. While Interview4 allowed job seekers to record interviews and share them for feedback, InterviewApp offers additional functionalities such as a training guide, the ability to review previous video recordings, and the opportunity to receive structured and elaborate feedback from peers and experts. Furthermore, the theoretical underpinning of using theories like SDT to design tools such as InterviewApp is missing in the previous literature.

2.3.3 Asynchronous Interview Systems

Online interviews for recruitment are typically conducted using computer-mediated communication like instant messaging, email, or video. Online interviews can be: synchronous, near-synchronous, and asynchronous [275]. Synchronous interviews happen in real time with simultaneous communication exchange. Near-synchronous interviews are near-immediate, ongoing posts and responses. With asynchronous interviews, there is a time-lapse between the communicating parties. Also called one-way interviews, they are usually conducted via online video interviews using internet-enabled digital devices. Asynchronous video interviews (AVI) and agent-based interviews have evolved as a tool to conduct the first round of screening as well as interview coaching. These are gaining increased attention due to their scalability and ease of use. In a recent article [195], the authors reviewed existing literature on AVIs and offered a framework for analysing the function of design elements in AVIs and their possible influence on the reactions and behaviours of applicants. Langer et al. [174] examined AVIs and videoconference (live) interviews and discovered that AVIs yielded higher participant ratings. In videoconference interviews and AVIs, participants judged fairness and chance to perform similarly. However, they thought that AVIs were creepier and less personal. In another experimental study [172], researchers analysed technologically sophisticated interview techniques by making the participants watch an interview with a virtual avatar. They discovered that AVI fairness is directly related to the probability of success, that transparent information is not always helpful, and that applicant reactions are unaffected by computer experience.

However, these research works focus on using AVIs for recruiting and not for training. In this work, we leverage the anytime, anywhere opportunity provided by AVI to build a self-help training platform supporting job seekers to enhance their chances at a job interview.

2.4 InterviewApp Design

InterviewApp's design is aimed at supporting the three BPNs posited by the SDT and, in turn, support job seekers' performance. The design goals were as follows: 1. To facilitate *autonomy* and practice with free will engagement; 2. To enable *competence* through comprehensive and informational feedback; 3. To foster *relatedness* through feedback from peer and expert. Additionally, we design InterviewApp to adhere to BST (i.e., instruct, model, practice, and feedback) methodology [176].

2.4.1 InterviewApp Features

InterviewApp was conceptualized as an asynchronous video interviewing and recording tool for training and feedback. It is available as a web application to job seekers. InterviewApp has three major user Personas: job seeker/interviewee, peer, and expert. We present below the main features of the InterviewApp, their potential role in supporting BPNs, and their link to BST.

2.4.1.1. Training Guide The Training Guide is a custom-made brief guide[6] that offers information on preparing for online interviews, general tips for successfully completing an interview, and different categories of interview questions with criteria to answer them. It suggests a few samples[7] for each type of question that the InterviewApp poses, along with guidelines for suitable responses. For example, the training guide explains the added value of answering in a structured manner using the STAR approach [21] of storytelling when answering past behaviour questions (e.g., "Tell me how you handled a difficult situation at work"). Every job seeker has instant access to the training guide from the header (Fig. 2.4e) on every page of InterviewApp.

Design Rationale. The training guide presents compelling rationales for why different categories of questions are regarded as vital in a job interview. This might help the job seekers reflect on the reasons for employing the tool and instill an interest in the actions they need to take to achieve their desired outcome, leading to volitional engagement and autonomy. We also suggest approaches for responding to those questions, although the optimal approach will ultimately be determined by the job seeker. Job seekers are free to use this manual as and when they like. Training guide aims at satisfying the instruct and model parts of the BST methodology. This is because it is designed to provide users with advice and recommendations for successfully completing an interview.

2.4.1.2. Record As a job seeker logs into InterviewApp, one of the main calls to action is to record a mock interview (Fig. 2.3a).[8] Job seekers had no preparation time after the questions were prompted on the screen and before the recording began, or in between questions. The time allotted for each question was a maximum of five minutes.

The content of the mock interview followed the usual structure of a job interview. Thus, each mock interview consisted of a total of five questions, prompted on the screen, with the first one always being an introductory question (e.g., "Tell me about yourself"). The other four were designed to assess the interviewee's competence and knowledge [131, 334]. These four questions were randomly picked from a pool of 137 questions[9] targeting six categories, i.e., background/work experience, leadership, personality, goals, skills, and behavioural & situational, which are frequently used in structured job interviews.[10]

Design Rationale. Although five minutes for each question is more than the AVI research recommendation [174, 195], we wanted to give job seekers greater control over their practice interview. They could take their time to think about their responses, make adjustments to their responses or delivery style, and ensure they expressed themselves effectively. Additionally, the interview questions are broad and open-ended, allowing job seekers to tailor their responses to fit their job description. This autonomy can allow job seekers to tailor the experience to their individual needs and preferences. The job seekers were able to make as

[6] See Sup. 1 for the complete training guide.

[7] Samples were inspired by `https://www.livecareer.com/resources/interviews`, last accessed August 2025.

[8] For clarity throughout the paper, mock interviews or mocks refer to practice interviews done on InterviewApp; job interviews refer to real-world employment interviews of the participants; qualitative interviews refer to the research interviews conducted by the authors of this paper.

[9] See Sup. 2 for the complete list of questions.

[10] see Appendix A.1 for question selection procedure

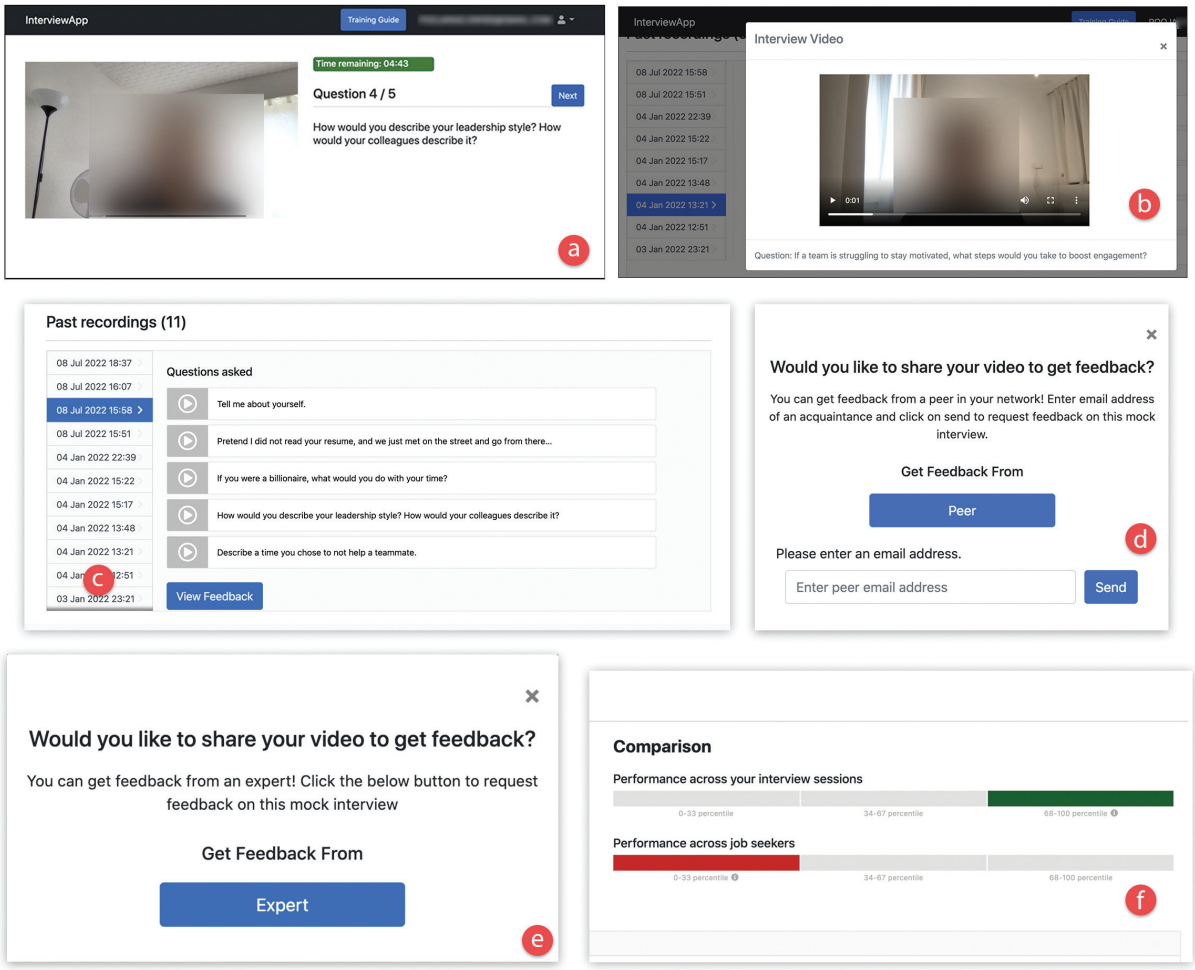


Figure 2.3: Features of InterviewApp: a) Mock interview recording with a question shown as a text prompt and a five-minute timer. b) Mock interview self-review. c) List of past mock video recordings. d) Request feedback from a peer. e) Request feedback from an expert. f) Comparison.

many attempts as desired to record the entire mock interview. The record feature fulfils the practice component of the BST.

2.4.1.3. Self-review / Playback This feature aims to facilitate self-review by providing the ability to rewatch the recorded mock interviews. The self-review is provided in three distinct views of the application. First, the job seekers can play the entire mock interview after recording. Second, on the home page, job seekers can browse a list of all the recorded mock interviews to date in chronological order (Fig. 2.3c). They can playback one video for each question, with the question's details underneath the video (as seen in Fig. 2.3b). Third, on the feedback page, we provide a choice to view their video while reviewing feedback. Additionally, the peer or the expert uses the playback feature to review and assess the interviewee's performance in the mock interview (see Sec. 2.4.1.4 for details).

Design Rationale. This feature allows job seekers to observe and analyse their performance, both verbal and non-verbal, in a reflective manner. This can help identify areas for enhancement and modify their performance, leading to iterative improvements and self-directed learning. Additionally, by providing access to the history of mock interviews, job seekers can reflect on performance progression over time. They may potentially link the video sequence with relevant factors like upcoming interviews, work-related stress, or personal issues, which could have impacted their performance. This information might help them become aware of different circumstances that can influence their behaviour and performance, further supporting competence [318].

2.4.1.4. Feedback The feedback feature has the following sub-features: request feedback, view feedback, view feedback details, and comparison.

(1) Request feedback: After each mock interview, job seekers have the option to request feedback through a pop-up modal (Fig. 2.3e) based on their group. For peer feedback, more details about the peer's email address are required (Fig. 2.3d). The peer might be any individual from the job seeker's network who is not involved in the study. A unique link to the page where an expert or a peer can rate the mock interview is created upon feedback request and shared as an email notification to the expert or peer. Additionally, the job seeker can request feedback from the home page on any of the past mocks.

(2) View feedback: Once a peer or an expert rates the mock interview, a notification is sent to the job seeker with a link to access the feedback (Fig. 2.4). This feedback can also be accessed from the home page for any of the past videos, if available. The performance of the job seeker is operationalized using a set of job attributes displayed on the feedback page (Fig. 2.4). The job interview performance attributes were chosen by consulting the literature in the field of personnel selection [147]. The list of these attributes with detailed descriptions[11] can be found in Table 2.1. The feedback is presented as a set of ratings from one to seven (overall and per attribute, Fig. 2.4b) and comments (Fig. 2.4c). The overall score is computed as an average of all the attribute scores. The expert or the peer can give overall comments and suggestions to the job seeker. For the expert feedback, these

[11]These job attributes serve a dual role as feedback for job seekers and as performance measures to assess improvements over the course of the study. See Sec. 2.5.3.1 and 2.6.1.4 for details.

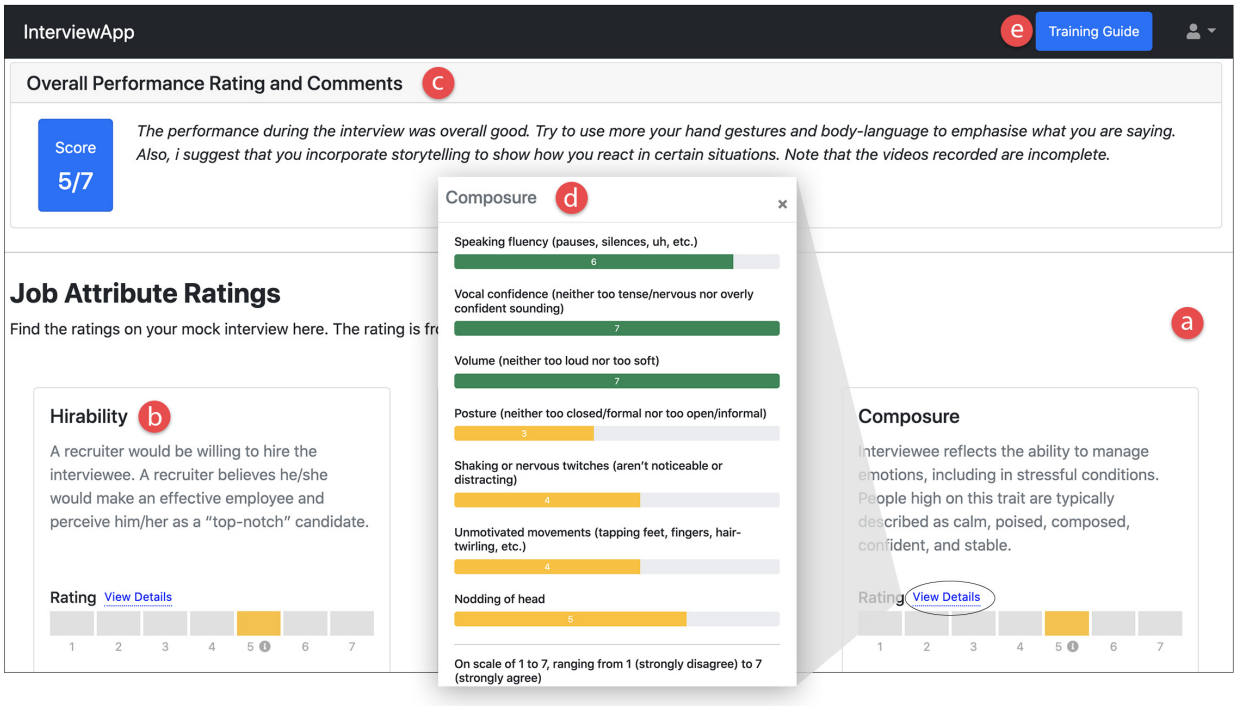


Figure 2.4: View Feedback: a) Feedback page. b) The seven job performance attribute ratings. c) Overall score and comments. d) Feedback detail view. e) Training guide.

comments mainly focus on the low score received. For example, if the job seeker receives a low score on the self-promotion, the comment will address that attribute with a one-line actionable suggestion – *"Try to take the opportunity to self-promote. Talk more about your achievements and give concrete examples to show your talent and demonstrate your potential value to the employer."*

(3) View feedback details: To help the job seekers gain more insights into the feedback and to make them aware of the specifics of the rating criteria, we incorporated the view details feature (Fig. 2.4d). Each job attribute is measured with a standardized scale (more details in Table 2.1). Job seekers can see the rating for each of the items composing the concerned rated attribute in the view details modal, appearing on the click of the view details link.

Design Rationale. The feedback page was designed to be structured and comprehensive, focusing on the important attributes usually assessed in a job interview. The overall comments from an expert or peer provide informational feedback with comments concerning the positive aspects of the performance and actionable suggestions on the negative aspects. This reinforces the job seekers' trust in their ability to succeed, promoting feelings of being relevant and competent [82, 83], potentially helping the job seekers improve performance, leading to competence satisfaction. Further, the ratings on the different job attributes, along with the detailed view, provide granular feedback focusing on the actionable elements that the job seekers can improve on.

The feedback feature was also designed to promote relatedness among job seekers through the factors of *acknowledgement* and *support*. Sharing their recorded video interviews with close peers or with an expert provides an opportunity to perceive *acknowledgement* and

receive *support*. Receiving the requested feedback can further augment this feeling. When a peer or expert views the video, it conveys a sense of being listened to and receiving someone's full attention.

Requesting and receiving feedback gives job seekers a sense of control over their learning process, allowing them to make decisions and take ownership of their progress. The feedback also presents opportunities for further action (engaging in more practice) to the job seekers, supporting autonomy. Further, as feedback aligns with the rationales presented in the training guide, it can increase job seekers' commitment to the training process. This alignment creates a sense of deeper agreement with the training and the path ahead to secure a job, increasing their volition and willingness to engage in the process.

The feedback also facilitates peers in providing structured feedback beyond open-ended comments, thus ensuring equitable opportunities for job seekers receiving peer feedback compared to those receiving expert feedback. Structured feedback, divided into discrete items and on a numeric scale, can improve the quality of feedback and elicit positive reviews from peers [138].

2.4.1.5. Additional autonomy-supportive characteristics The inherent *anytime, anywhere* nature of AVIs gives job seekers the flexibility to practise as and when they deem fit. InterviewApp allows them to explore any direction rather than forcing them to a linear path, fostering a sense of ownership over the preparation process. To this end, job seekers can determine when and how to engage with InterviewApp, such as recording mocks, listening to their recordings, and requesting feedback on new and old mocks in any sequence they prefer and as many times as they wish.

2.5 Method

To evaluate the effect of InterviewApp, we conducted a four-month long study. The experiment was conducted with a mixed-design setup. The three experimental groups (*self*, *peer*, *expert*) and *control* form the between-subject factor. The study consisted of three phases: pre-treatment, four months of treatment, and post-treatment. Outcome variables were measured pre- and post-treatment, making time the within-subject factor. Fig. 2.5 represents the phases of the study and the different groups with their access to features on InterviewApp. This study was pre-registered.[12]

2.5.1 Participant Recruitment and Characteristics

We recruited the participants via the online recruitment platform *Prolific*.[13] We conducted a power analysis to determine the required sample size. To reach an effect size (f) of 0.40 with alpha=0.05 and power=0.95, a total of 112 participants were required. Taking into account dropouts, technical issues with mock interview videos, and optimizing for the budget, we chose to recruit 228 participants. We used the following screening criteria: (i) actively

[12]Pre-registration details: `https://aspredicted.org/7M4_L8R`.

[13]See `https://www.prolific.co/`, last accessed August 2025.

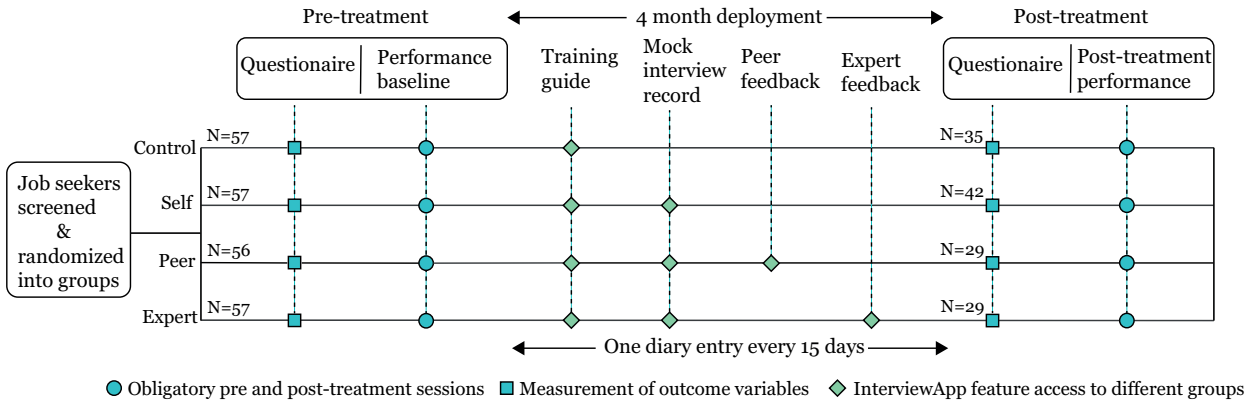


Figure 2.5: The different phases of our study. The figure presents the number of participants randomly assigned to each group and the number of participants who completed the post-treatment. The four-month deployment phase represents the differences between the groups' access to the features of InterviewApp.

seeking employment; (ii) unemployed, partly employed or not in paid work at the start of the study; (iii) 18 years of age or older; (iv) willing to record a video; (v) access to a webcam. We restricted the participants to be located in the United Kingdom and Ireland. This was to ensure that the unemployment and the job landscape were the same among all participants. These criteria provided us with a pool of approximately 1,250 participants. The recruitment process started in September 2021 and ended in December 2021, providing 228 participants.

2.5.1.1. Experimental Groups After agreeing to the informed consent form, these participants were randomly assigned to one of the three experimental conditions or the control, with 57 participants in each group. Participants were not informed of their designated groups. Participants in the control group (henceforth *control*) had access to only the training guide feature. Participants in the self-review group (henceforth *self*) had access to the guide, record, and self-review features. Participants in the peer group (henceforth *peer*) had access to the guide, record, and self-review features. Finally, each recorded interview had the option of receiving feedback from peers. Participants in the expert group (henceforth *expert*) had access to the guide, record, and self-review features like the peer and self groups. Each recorded interview had an option to receive feedback from an expert. We provided them with an expert who was trained to evaluate interviews. This was the same rater who rated all the pre- and post-treatment videos (more details in Sec. 2.5.2).

2.5.1.2. Participant Characteristics. Of the 228 participants, one from *peer* did not complete phase 2 of the pre-treatment session, effectively making the number of participants 227. Of the 227 participants, all were actively seeking a job, comprising 143 females and 84 males. The age distribution of participants was as follows: 18-29 ($n = 123$), 30-39 ($n = 58$), 40-49 ($n = 33$), 50-59 ($n = 9$), 60+ ($n = 2$), with a mean of 30.7 (SD = 10.0).

After four months, 135 participants responded to the post-treatment session, representing a 40.5% attrition rate. This attrition rate aligns with previous studies on digital unemployment tools [94, 96] and longitudinal research conducted using Prolific [168]. We considered these 135 participants for further analysis. The number of participants retained in each

group is: 35 (61%) in *control*, 42 (74%) in *self*, 29 (52%) in *peer*, and 29 (51%) in *expert*. Among the 135 participants, 91 were female, and 44 were male, with a mean age of 30.9 (SD = 9.6). 48% (n = 65) were part-time employed, 22% (n = 29) working full-time, 19% (n = 25) unemployed, 7% (n = 9) reported other (about to begin as a student), 2% (n = 3) self-employed, 2% (n = 3) due to start a new job within the next month and one participant not in paid work. 56% (n = 76) of 135 were actively seeking a job at the time of the post-treatment. Each participant appeared for, on average, 1.99 interviews during the study period.[14] Among the 18 qualitative interview participants, 12 were female, and six were male, with a mean age of 36.9 (SD = 10.26); 13 were part-time employed, three were full-time employed, one was a student, and one was unemployed. 12 participants were actively seeking a job at the time of the interviews.

2.5.2 Procedure

2.5.2.1. Ethics. The study was approved by the authors' Institutional Review Board (IRB). All participants were asked to provide informed consent.[15] We compensated the participants at an hourly rate of GBP 7.5 for the pre- and post-treatment sessions and GBP 5 for each diary entry. The pre- and post-treatment sessions could last for a maximum of 30 minutes each, and each diary entry could last for a maximum of 2 minutes. Participants were paid an incentive of 20 GBP for the qualitative interviews, which lasted approximately an hour. Prolific anonymizes participants using a unique participant ID (24-character alphanumeric). All communication with the participants took place exclusively through their anonymized internal messaging service. The research materials for this study were shared in accordance with the criteria for transparency and research ethics outlined by Salehzadeh Niksirat et al. [274].

2.5.2.2. Pre-treatment Session. We conducted the pre-treatment sessions in two phases. *Phase 1* included a questionnaire asking about participant demographic information, employment, job search status and questions from the BPN Satisfaction and Frustration scale and Generalized Self-efficacy scale (GSE) (more details in Sec. 2.5.3.1). In *phase 2*, participants were asked to give a mock interview using InterviewApp. These were rated by the expert on seven performance measures (see Table 2.1) to determine the interview performance baseline of participants for measuring the change in performance over the course of the study.

2.5.2.3. Four-month Deployment. After the pre-treatment session, the job seekers were informed that InterviewApp would be available for use free of charge, as many times as they needed, for a period of four months. During that period, they received timely reminders (every 15 days) that they could use InterviewApp for practice and, depending on the group, instructions on how to request feedback with an example feedback page. They also received messages upon the availability of the feedback with instructions to view the feedback.

[14]This is calculated using the self-reported job interviews from the diary entries. This is an approximate measure, as there can be overlap in the timeframe between the diary entries.

[15]The consent form used can be found in Sup. 3 .

2.5.2.4. Post-treatment Session. The post-treatment session was similar to the pre-treatment session. In Phase 1, questions regarding the number of actual job interviews they attended and real-world job offers in the four-month period were asked, along with the BPN and GSE scales. In Phase 2, we did an assessment on a mock interview to determine the post-treatment interview performance to compute the difference in job seekers' performance (i.e., performance measures, see Table 2.1).

With the current study design, we probe the effect of training on interview motivation and performance. The circular effect of motivation on training cannot be probed with this design, as we did not record the engagement of job seekers with InterviewApp after the post-treatment (nor the motivation of the job seeker at different points throughout the experiment).

2.5.2.5. Diary Maintenance. To understand the job seekers' interview-related activities outside the InterviewApp, we conducted a diary study to complement the original study. This was spread across the four-month period of the study with six diary entries at equally spaced time intervals of two weeks. Here, we asked questions like their job interviews in the past two weeks, any interview-related practice, and feedback outside of InterviewApp. [16]

2.5.2.6. Qualitative Interviews. We conducted semi-structured qualitative interviews[17] to gain insights into job seekers' perceptions of online interviews and training, InterviewApp, and the feedback they received using the platform. We recruited participants following the post-treatment sessions. Stratified sampling was employed based on outcome variables, number of mock interview videos, and feedback requests (for the peer and expert groups), ensuring representation of the entire InterviewApp user population. Fifteen participants were chosen from each of the experimental groups, along with five participants from the control group, resulting in a total of 50 participants from different strata. In total, 18 interviews were conducted, with six participants each from the peer and expert groups, five from the self, and one from the control. Further interviews were discontinued after reaching saturation [53] and observing minimal response variations.

2.5.2.7. Expert Rating Procedure The expert rated the mock interviews following a procedure that involved four steps. (i) Training: A rating protocol was established, and three independent raters, all from the Organisational Behaviour department of the university, familiarized themselves with it. (ii) Calibration: Initial calibration involved three raters assessing five videos, resolving any score discrepancies through discussion. Once agreement was reached between the three raters, two of them coded a subset of videos to assess reliability. (iii) Inter-rater agreement: Two raters evaluated 15 pre-treatment videos, yielding strong Intraclass Correlation Coefficients values (ICC) [167]: 0.90 for hireability, 0.87 for communication, 0.89 for positive impression, 0.81 for self-promotion, and 0.89 for confidence. Composure and storytelling yielded fair ICC values of 0.66 and 0.57, respectively. (iv) One rater from the previous step (designated as the expert) rated the rest of the pre-

[16]The diary template can be found in Sup. 4 .
[17]The complete protocol can be found in Sup. 5 .

Table 2.1: Details of the outcome variables used as dependent and mediating variables in the data analysis.

Type of Measure	Outcome variable	ID	Self/other reported	Source questionnaire	No. of items	Description
BPN measures	Autonomy	O_1	Self	BPN Satisfaction and Frustration Scales [66]	8	Feeling in control of his/her behaviour, experiencing volition [271].
	Competence	O_2	Self	BPN Satisfaction and Frustration Scales [66]	8	Feeling a sense of mastery, skill and effectance [271].
	Relatedness	O_3	Self	BPN Satisfaction and Frustration Scales [66]	8	Feeling a sense of belonging and connection [271].
Self-efficacy	Self-efficacy	O_4	Self	Generalized Self-Efficacy scale [283]	10	Judgement in one's ability to perform well in an interview [20].
Anxiety	Anxiety	O_5	Other	Self-developed	1	How stressed and anxious the interviewees presented themselves during the interview [146].
Performance measures	Hireability	O_6	Other	Previous research on hireablity [202]	3	Likelihood of an individual getting hired, measure the potential that a job seeker will be an effective employee [202]
	Communication	O_7	Other	Conversational Skills Rating Scale [294]	5	Ability to articulate (and receive) ideas and information clearly, accurately, and convincingly [147].
	Composure	O_8	Other	Conversational Skills Rating Scale [294]	7	Regulation and management of one's emotions in stressful conditions. Described as calm, poised, appear at ease and comfortable [147].
	Positive impression	O_9	Other	Self-developed	3	Lasting impact an individual has after the interview, influenced by non-verbal behavior [151].
	Self-promotion	O_{10}	Other	Previous research on impression management [41]	4	Impression management tactics; refers to an attempt to evoke favourable character perceptions about oneself to elicit respect for one's abilities rather than just being liked [123, 326].
	Confidence	O_{11}	Other	Self-developed	3	How confident the interviewees presented themselves during the interview [147].
	Storytelling	O_{12}	Other	Self-developed	1	Answering past behaviour questions with a narrative, providing information on the situation, task, action and results [21].

and post-treatment videos, along with other videos recorded by the participants during the study, and for which they requested feedback from an expert. Note that the expert rater was blind to the experimental groups and the feedback history of the participants. Furthermore, due to the study timeline (i.e., Sept 2021 to Apr 2022), some of the ratings of the pre-treatment, post-treatment, and those produced during the study overlapped.

2.5.3 Measures

2.5.3.1. Outcome Variables In our experiment, we considered the three BPNs, interview self-efficacy, interview anxiety, and interviewee performance as the outcome variables. Table 2.1 gives the details about the 12 outcome variables and the standardized scale used to measure each of them. For uniformity, all the outcome variables were measured on a seven-point Likert scale.[18] The items were averaged to compute one composite score for each outcome variable. For the BPN measures, we followed Chen et al. [65] in combining the need satisfaction and reversed need frustration items to calculate a score, and finally averaged them to compute a composite score for each need.

2.5.3.2. Process Variables We logged various interactions of the participant with the app throughout the four-month study period. With the help of these logs, the following data was computed per participant: (a) number of logins, (b) number of mock interview videos, (c) number of video playbacks, (d) number of guide views, (e) duration of the mock videos, (f) cumulative time spent on the app, (g) number of feedback requests, (h) number of feedback views, (i) number of feedback detail views, (j) number of views with feedback not yet available, (k) number of attempted feedback views without request. The variables from

[18]See Sup. 6 for the complete questionnaire used to measure all the outcome variables.

(a) to (f) are applicable to all groups, whereas the variables from (g) to (k) are applicable only to *peer* and *expert* groups.

2.5.4 Data Analysis

The quantitative dataset[19] used to perform statistical tests included the outcome variables mentioned in Table 2.1 along with the following: *(i)* self-reported number of job offers in the period of the study from the post-treatment session, *(ii)* self-reported number of real job interviews from the diary entry every 15 days, *(iii)* process variables (Sec. 2.5.3.2).

Multiple statistical tests were performed. Data blinding was carried out by scrambling the group labels before the analysis.[20] For each model presented below, the results are reported in Sec. 2.6. All the models were fitted using the lm or the lme4 and lmerTest [26] packages of the statistical software R. We used the difference between the pre- and the post-phases ($\Delta = post - pre$) as the predictor for each of the outcome variables in all the tests.

To analyze the qualitative interview data, we followed the thematic analysis approach [45, 46]. The entire procedure was carried out by the first author. An inductive approach to thematic analysis was employed primarily, with codes and themes derived from data content. First, the data was familiarized, followed by recursive coding, in which codes were revisited and refined as the coding process progressed. This procedure yielded 235 codes. A thematic map was created using the codes, which were then grouped into 31 sub-themes to indicate their prevalence throughout the data set. This process ultimately produced five main themes.

An additional coder, thoroughly briefed on the inductive coding procedure, independently coded a representative 15% subset of the interviews (three out of 18 interviews, one randomly selected from each of self, peer, and expert groups), identifying a smaller codebook. This was done to enhance interpretation and reflexivity rather than to reach a consensus about data coding [47, 292]. The two coders discussed the codebook and reached inter-coder reliability of 0.84 using the equation proposed by Miles et al. [211]. However, we opted for a single coder based on resource availability and the first coder's in-depth familiarity with the data.

2.5.4.1. Confirmatory Analysis. To test our hypotheses, we built separate multiple regression models to predict each outcome variable (Table 2.1) with the groups (categorical: control, self, peer, and expert) as the predictor variables and age (numeric) and gender (categorical: M and F) as control variables. This gave us three models for the BPN measures, one for anxiety, one for self-efficacy, and seven models for performance measures. All pairwise comparisons were done with Tukey's adjusted p-values [308].

2.5.4.2. Engagement Analysis. To determine the effect of engagement data on the outcome variables, we performed multiple regression with process variables predicting the outcome variables. We used variables (numeric) except those related to feedback (i.e., except (g) to (k) from Sec. 2.5.3.2) as predictors and all outcome variables as dependent variables

[19]The details on the dataset, as well as the data analysis protocol, are available in Sup. 7 .

[20]R workbooks with results of blinded and unblinded analysis can be found in Sup. 8 .

to build models with data from all groups. We perform a data-driven subgroup analysis with only peer and expert groups, with all the process variables as predictors. In all the regression models for the engagement analysis, the factors were selected based on the AIC procedure.

2.5.4.3. Mediation Analysis. We considered the experimental groups as the independent variables (IV), the three BPN measures (numeric), self-efficacy (numeric), and anxiety (numeric) as the mediators, and each of the seven performance measures in Table 2.1 as the dependent variables (DV). The models to check for the effect of IV on the mediators were described in Sec. 2.5.4.1. We built additional multiple regression models to check the effect of mediators on the performance measures while controlling for the IV along with age and gender.

2.5.4.4. Additional Analysis. We also wanted to understand if the InterviewApp intervention had any real-world impact. We performed an exploratory analysis by developing a logistic regression model to understand the effect of the InterviewApp engagement on the participants' actual job offers (as a proxy to interviewer ratings in Fig. 2.1). The job offer information was converted into a binary variable indicating if participants were offered or changed jobs. The IVs were process variables except those related to feedback with all groups and all process variables for the peer and expert groups.

To gain better insights into the factors that influence job seekers' real-world job interviews, a linear mixed-effect regression (LMER) model was fitted. IV included the process variables (i.e., (a) to (f) from Sec. 2.5.3.2) except those related to feedback. As the information on the number of job interviews was obtained through diary entries, we first estimated the engagement data for every 15 days for each participant. Then, the regression analysis was run with the number of job interviews as the dependent variable. A random intercept accounted for the participant effect.

2.6 Results

Fig. 2.6 presents a visual summary of the results of the statistical analysis. Overall, our results show that the InterviewApp intervention had a significant effect on the job seekers' autonomy. Furthermore, we found that autonomy is a mediator of performance. The request feedback feature had a positive impact on autonomy and relatedness. The duration of the mock interview had a positive effect on most of the performance measures. The feedback page views and the detail views had a significant effect on communication, composure, self-promotion, and confidence. We will review these results in detail below. We will focus on and present the details of statistically significant results in order to be succinct and clear, and we will only mention non-significant outcomes that are pertinent.[21] We describe the results of the statistical tests in the same order as the analysis in Sec. 2.5.4. Finally, we present the results of the qualitative analysis, where we describe the five salient themes generated.

[21]R workbooks including all analyses and detailed output are available Sup. 8 .

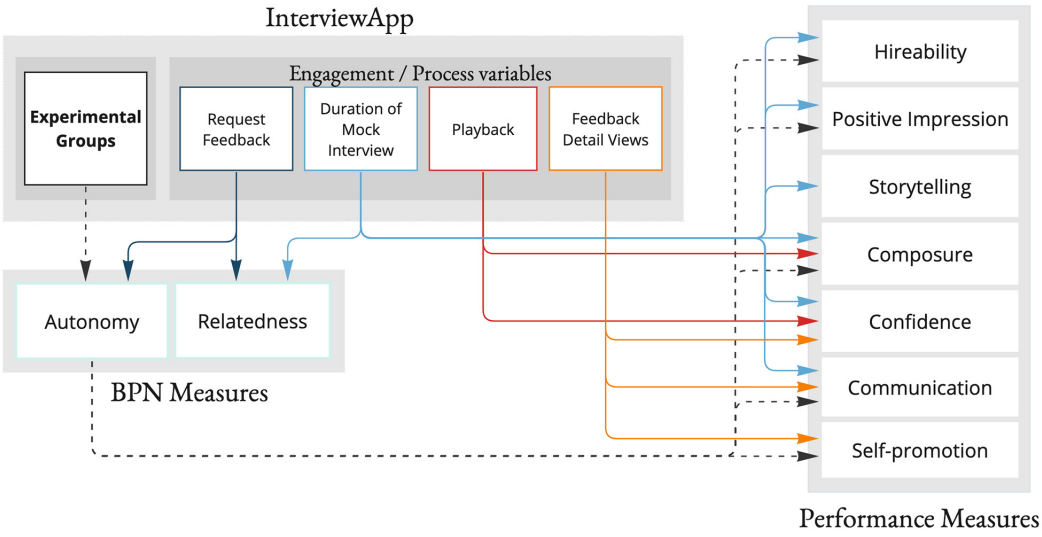


Figure 2.6: Visual summary of the quantitative results. Lines represent the significant effect InterviewApp intervention and job seekers' engagement had on the outcome variables tested in the study. Only those process variables that had a positive significant effect on more than one outcome variable are depicted. The dashed lines represent the mediation effect. Request feedback and feedback details views are applicable only to expert and peer groups.

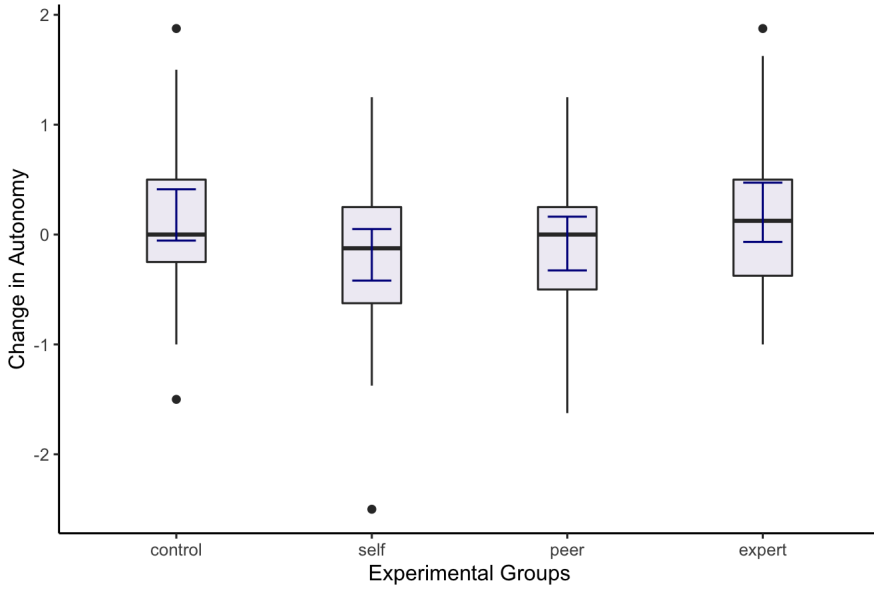


Figure 2.7: Change in autonomy (Δ) from pre-treatment to post-treatment across experimental groups. The error bar indicates a confidence interval of 95%.

2.6.1 Effect on Outcome Variables

2.6.1.1. BPN Measures (*H1, H2*) The effect of InterviewApp through the experimental groups was significant on the autonomy[22]construct only ($F_{(5,129)} = 3.35, p < .01, R^2_{adj} = 0.08$);

[22]All the outcome variables tested throughout this section refer to the difference between the pre and the post-phases ($\Delta = post - pre$).

Table 2.2: Coefficients of the regression models built with engagement data from all groups predicting the outcome variables O_6: hireablity, O_7: communication, O_8: composure, O_9: positive impression, O_{11}: confidence. Column X with (a) to (k) indicates the respective independent variables described in Sec. 2.5.3.2. Only significant models with independent variables that remain after selection are shown here. *p<0.05; **p<0.01; ***p<0.001

X	O_6	O_7	O_8	O_9	O_{11}
Videos (b)	-0.378^{*} t = −2.266	-0.370^{*} t = −2.289	−0.269 t = −1.864	-0.495^{**} t = −2.657	-0.478^{*} t = −2.154
Playbacks (c)	0.079 t = 1.736	0.064 t = 1.454	0.097^{*} t = 2.458	0.096 t = 1.898	0.121^{*} t = 2.001
Duration (e)	0.001^{***} t = 3.971	0.001^{***} t = 4.445	0.0004 t = 1.941	0.001^{***} t = 3.546	0.001^{*} t = 2.237
Intercept	−0.112 t = −0.316	−0.162 t = −0.471	0.041 t = 0.133	0.127 t = 0.320	0.602 t = 1.277
Adjusted R^2	0.102	0.120	0.046	0.088	0.043
F Statistic (df = 3; 131)	6.067^{***}	7.114^{***}	3.154^{*}	5.329^{**}	3.009^{*}

the difference between the expert group and self group was significant ($p < .05$). This finding reveals that the experimental groups were associated with the change in autonomy (which we did not hypothesize), and the expert group showed an increased autonomy compared to the self group. On the other hand, hypotheses *H1* and *H2* on competence and relatedness were not confirmed ($p = 0.55$ and $p = 0.69$ respectively). Although the feature requesting feedback had an impact on relatedness (see Sec. 2.6.1.2), an overall effect of the experimental manipulation was not observed. Some plausible explanations for this could be *(i)* the absence of feedback to most participants in the peer group who asked for it, despite the expectation that someone would be available to provide it; *(ii)* participants wished to receive extensive feedback, for example, comments to improve each job attribute, to have an effect on their competence. The themes of *need for relatedness* and *criteria for competence* (Sec. 2.6.5.3 and 2.6.5.4) explain these and more reasons in detail.

2.6.1.2. Effect of Engagement with InterviewApp (H3) The results of the engagement analysis can be seen in Table 2.2. This table describes only the regression models where there was a significant effect on the outcome variables.[23] The performance measures for hireablity, communication, positive impression, and confidence were affected by the quantity and duration of mock interview videos. While duration was positively associated, the number of videos was negatively associated with these performance measures. This implies that the change in performance is positively linked with participants providing longer answers to the questions than when they recorded more mock interviews. The playback feature had a significant effect on the composure and confidence measures. Composure and confidence are directly linked to candidate appearance [147]. Despite the fact that participants did not like this feature (see Sec. 2.6.5.5 for details), they may have used playback to self-reflect and increase their composure and confidence.

[23]Refer Sup. 9 for the table with complete results.

Table 2.3: Coefficients of the regression models built with engagement data from peer and expert groups predicting the outcome variables O_1: autonomy, O_3: relatedness, O_6: hireablity, O_7: communication, O_8: composure, O_9: positive impression, O_{10}: self-promotion, O_{11}: confidence, O_{12}: storytelling. Column X with (a) to (k) indicates the respective independent variables described in Sec. 2.5.3.2. Only significant models with independent variables that remain after selection in more than one model are shown here. *p<0.05; **p<0.01; ***p<0.001

X	O_1	O_3	O_6	O_7	O_8	O_9	O_{10}	O_{11}	O_{12}
Logins (a)	−0.062 t = −1.533	−0.078 t = −1.617		−0.186* t = −2.029					
Duration (e)	−0.0003 t = −1.624	0.0004* t = 2.106	0.001*** t = 3.536	0.001** t = 2.853	0.001** t = 2.805	0.001* t = 2.307	0.001 t = 1.930		0.001* t = 2.400
Request (g) Feedback	0.241** t = 3.085	0.215* t = 2.323							
No Feedback (j) Views	−0.181 t = −1.761	−0.242 t = −1.985					−0.406 t = −1.670		
Feedback (h) Views			0.305 t = 1.684	0.314 t = 1.483	0.485*** t = 3.628				
Feedback (i) Detail Views				0.293* t = 2.462		0.172 t = 1.522	0.251* t = 2.123	0.261* t = 2.372	
Intercept	0.311 t = 1.792	−0.098 t = −0.476	−0.465 t = −0.713	−0.636 t = −1.550	−1.013*** t = −4.113	−1.160** t = −3.106	−0.811* t = −2.078	−0.026 t = −0.106	0.139 t = 0.414
Adjusted R^2	0.180	0.113	0.195	0.308	0.283	0.169	0.171	0.075	0.123
F Statistic	4.122** (df = 4; 53)	2.812* (df = 4; 53)	5.614** (df = 3; 54)	7.330*** (df = 4; 53)	12.249*** (df = 2; 55)	6.784** (df = 2; 55)	4.932** (df = 3; 54)	5.626* (df = 1; 56)	4.989* (df = 2; 55)

Sub-group Analysis Table 2.3 presents the results of the sub-group analysis with data from peer and expert groups. The engagement data had a significant effect on all outcome variables except competence, self-efficacy, and anxiety. For conciseness, we will review only the significant factors while focusing on the feedback-related features.[23] The number of feedback requests was positively associated with the change in autonomy and relatedness. These results support *H3* only for the peer and expert groups. This suggests that the self-determined decision to ask for feedback from peers or from an expert promoted autonomy and relatedness in the job interview training process. The number of feedback detail views was positively linked with communication, self-promotion, and confidence. The participants may have been able to incorporate the feedback elements into their interviews with the help of the information provided in the detailed views. The overall feedback page view impacted the composure scores. An explanation for this could be that with more feedback views, as reflected by the page view count, participants might have gained a broader perspective on their performance, assuring them of their abilities. Consequently, this could have regulated their composure.

2.6.1.3. Self-efficacy (*H4*) and Anxiety The models to predict self-efficacy and anxiety ($p = 0.78$ and $p = 0.15$) were not significant. Our hypothesis (*H4*) that practice with self-review and feedback on mock interviews would lead to increased self-efficacy was not validated. The qualitative interviews investigate these findings. Theme 5 in Sec. 2.6.5.5 sheds light on the factors contributing to the infrequent use of self-review by the participants. These factors could potentially explain the lack of impact on self-efficacy.

2.6.1.4. Performance Measures (*H5*) None of the seven regression models built for the performance measures was significant. *H5* was not confirmed. In spite of significant positive associations observed between various features of InterviewApp and performance measures–

as outlined in Sec. 2.6.1.2, the hypothesized impact of the experimental intervention on performance measures was not found. Our qualitative analysis suggests that several factors may have contributed to this outcome, including the need for more comprehensive feedback (see Sec. 2.6.5.3), and the influence of variables external to our experiment (see Sec. 2.6.5.1), which could have overcrowded the effect of the experimental manipulation on performance. We acknowledge this as a limitation of the current results that will have to be verified by further research (see Sec. 2.7.4). However, the qualitative interviews also suggest that the job seekers found the feedback to be actionable and suggested areas of improvement. Though the models predicting storytelling and self-promotion ($p = 0.37$ and $p = 0.32$) were not significant, several participants liked the storytelling and the self-promotion aspects of the feedback. They noted that it gave them a distinctive takeaway for upcoming job interviews. *"Storytelling, presence of narrative content, it's something that you never think about, but it's nice when you know how to tell stories to the interviewer."* [F, self, 35].[24] We believe that compared to other metrics (e.g., hireability), storytelling and self-promotion are more concrete and actionable, and hence were chosen specifically above the others.

2.6.2 Mediation Effect on Interview Performance (*H6*)

The mediation analysis results are outlined in Fig. 2.8 and Table 2.4. Mediation analysis was performed with autonomy, competence, relatedness, self-efficacy, and anxiety as mediators (M_x) on each of the seven performance measures (Y), with the experimental groups as independent variables (X). The results of the model to check the effect of X on M_x have already been described in Sec. 2.6.1.1 and Sec. 2.6.1.3. To understand the indirect effect, we examined the effect of M_x on Y. From Table 2.4, we can see that the InterviewApp intervention had several significant indirect effects on performance measures. A significant indirect effect was seen on hireablity, communication, composure, positive impression, and self-promotion. As X had a significant effect on autonomy (M_1) only (described in Sec. 2.6.1.1), we can say that autonomy is the mediator. Therefore, we conclude that the experimental groups of the InterviewApp intervention indirectly affected the performance via autonomy. [25] As we did not hypothesize the direction of the effect in H6, these results support *H6* only for autonomy.

However, for each performance measure, the coefficient of autonomy (b_1) is negative when significant; meaning that an increase in autonomy is negatively associated with an increase in performance measures. The potential impact of extrinsic factors on training could explain this finding. The qualitative interviews support this explanation. The upcoming job interview schedule or the fortuitous availability of InterviewApp when a large majority of job interviews were conducted online may compel performance improvement brought on by training rather than allowing for free will. As we will discuss later, the aforementioned rea-

[24] We refer to the participants using their gender, group, and age, respectively.

[25] Note that the total effect of the groups (X) on performance measures (Y) is not significant (as seen in Sec. 2.6.1.4). However, mediation no longer imposes this precondition: as stated in Bollen [42], "A lack of correlation does not disprove causation. It is only when we isolate the cause and effect from all other influences that correlation is a necessary condition of causation." [42, p. 52] and in Hayes [133] "Mediation analysis as practised now no longer imposes evidence of simple association between X and Y as a precondition." [133, p. 80].

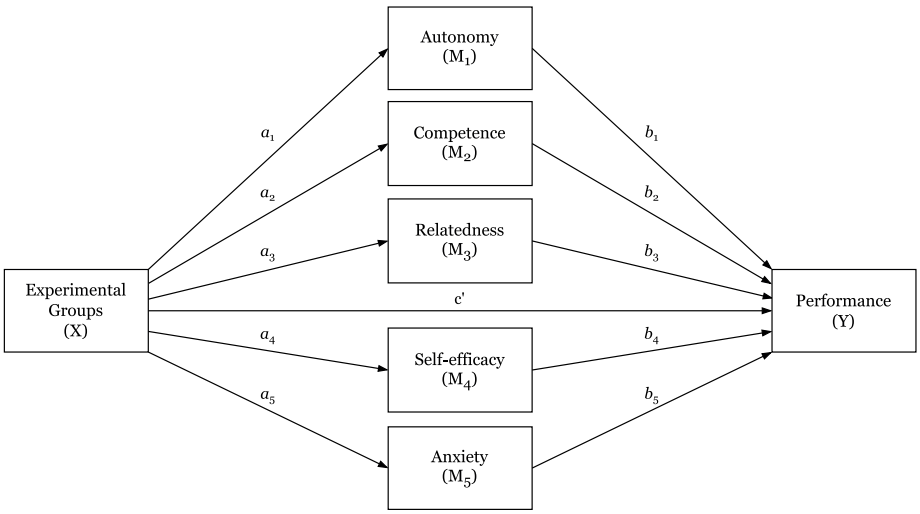


Figure 2.8: Mediation model.

Table 2.4: Mediation results with experimental groups (X), autonomy, competence, relatedness, self-efficacy, anxiety as mediators (M_x) and each of the seven outcome variables (Y). Regression coefficients are $a_x(X \rightarrow M_x)$, $b_x(X \longrightarrow M_x)$, c' (direct $X \longrightarrow Y$), c (total $X \longrightarrow Y$). Note that X is a categorical variable with four factor levels. For all a_x, c', c, three co-efficients for three factor levels of self, peer and expert are shown; control is the reference level. Significant results are bold. *p<0.05; **p<0.01; ***p<0.001.

M_x	Autonomy		Competence		Relatedness		Self-efficacy		Anxiety	
Y	a1	b1	a2	b2	a3	b3	a4	b4	a5	b5
Hireability		**-0.482***		0.039		**0.346***		-0.060		**-0.156***
Communication		**-0.431***		-0.035		0.248		-0.205		**-0.218*****
Composure	**(-0.345*,**	**-0.362****	(-0.028,	-0.072	(0.006,	0.103	(0.042,	-0.070	(0.786,	**-0.193*****
Positive impression	**-0.261,**	**-0.603****	0.031,	0.105	-0.063,	0.396	0.244,	-0.093	0.732,	**-0.189****
Self-promotion	**0.102)**	**-0.726****	0.059)	0.032	0.182)	0.445	0.126)	-0.149	-0.011)	**-0.306*****
Confidence		-0.076		-0.132		0.178		-0.278		**-0.422*****
Storytelling		-0.347		-0.154		0.119		0.026		**-0.204****

	Direct Effect	Total Effect
Y	c′	c
Hireability	-0.167, - 0.191, -0.055	-0.125, -0.216, -0.045
Communication	-0.08, -0.033, 0.264	-0.109, -0.148, 0.240
Composure	0.021, -0.319, 0.378	-0.007, -0.392, 0.350
Positive impression	-0.262, -0.330, -0.197	-0.207, -0.356, -0.190
Self-promotion	0.132, 0.722, 0.212	0.138, 0.624, 0.205
Confidence	-0.166, -0.209, -0.067	-0.479, -0.583, -0.081
Storytelling	-0.185, 0.125, 0.620	-0.219, 0.061, 0.603

sons may have been a few of the numerous extrinsic influences putting pressure on individuals to train, restricting their autonomy (see Sec. 2.6.5.1 and 2.7.1).

Anxiety was not found to be a mediator in our framework between the experimental manipulations and the performance (see Table 2.4). However, anxiety is significantly associated to all seven performance measures with a negative coefficient (b_5). This result is in line with prior literature [72, 206], as anxiety is negatively linked to performance. An anxious interviewee will tend to perform less well on all job attributes. One of the participants explained how their interview anxiety and nervousness can be a source of frustration. *"I know how to do all my jobs, I could do them in an eye shot, back to front, you name it. But to try to think of scenarios and things, I just totally go blank"* [F, expert, 37].

2.6.3 Additional Analysis: Influence on Real-World Job Activities

First, we looked at whether the job offers received by the participants during the experiment were influenced by the intervention. The logistic regression model used to predict the presence of a job offer did not show significant results. One employment offer was made to at least 78 of the 135 participants. It must be noted that this metric could be imprecise because employment offers were gathered from the participants at the post-treatment; it may have been impacted by recall bias. Beyond the purview of this intervention, other outside variables, aside from participant performance, might affect a job offer.

Next, we investigated whether the use of InterviewApp had any impact on job seekers' real-world job interviews to alleviate the above problem. From the LMER model (Sec. 2.5.4.4), we can deduce that there is a significant association between engagement data and the number of real-world job interviews. The model shows a linear decrease in the number of job interviews with an increase in engagement with InterviewApp for each period of 15 days. The factors, number of logins ($\beta = -0.0587, t_{129} = -2.167, p < 0.05$) and videos ($\beta = -0.2140, t_{129} = -4.533, p < 0.001$) show a significant effect. There is a time delay between when the participants practice and when the job interviews take place, indicating that they train well in advance. The findings from the qualitative analysis support this. See Sec. 2.6.5.3 for more details.

2.6.4 InterviewApp Engagement: Descriptive Statistics

A total of 6273 interactions were logged with InterviewApp, including the data from participants who dropped out of the study. Participants generated a total of 440 mock interview videos. 135 participants who completed the post-treatment session of the study generated a total of 318 mock videos and 1580 minutes of recording with a mean of $M = 2.4(SD = 0.8)$ videos and $M = 11.7(SD = 8.02)$ minutes per participant. There were a total of 457 logins ($M = 3.4, SD = 2.3$) and 167 playbacks ($M = 1.2, SD = 2.7$).

There were 63 total requests for feedback, with 50 from expert ($M = 1.7, SD = 1.3$) and 13 from peer groups ($M = 0.4, SD = 0.9$). Further, there were 32 feedback views ($M = 1.1, SD = 1.2$) and 34 detailed feedback views ($M = 1.2, SD = 2.5$) from the expert group. Participants had entered their own email addresses in 61%($n = 8$) of the 13 peer requests. Only one participant out of the remaining five received feedback and made two feedback views and eight detailed feedback views. In the qualitative interviews, we investigated the reason for these low feedback numbers in the peer group.

A few peer group participants from the qualitative interviews figured someone would be on standby on InterviewApp to offer feedback, so they entered their email addresses. This can be attributed to a usability issue of InterviewApp. The remaining participants made the same assumption, and when they realised that the video was to be shared with a peer, they did not request it. *"It might have been a good feature to have someone who you don't know reviewing your questions."* [F, peer, 41]. Sec. 2.6.5.4 discusses some of the reasons for low peer feedback.

2.6.5 Qualitative Interview Results

We present the five main themes of the interview findings. For consistency, we use the following determiners-to-percentage mapping based on the frequency: a few for n = 1-3 (0-20%), several for n = 4-7 (20-40%), about half for n = 8-11 (40-60%), most for n = 12-15 (60-80%), and almost all for n = 16-18 (80-100%) participants.

2.6.5.1. Theme 1: External variables influencing interview training and performance This theme elaborates on the findings in Sec. 2.6.2, identifying external factors affecting job seeker training and performance. These could be divided into factors *outside* of and *within* the job seeker's control. These are collectively categorized as extrinsic to job seeker preparedness and training with InterviewApp.

About half of the participants found **InterviewApp's launch timely** due to the prevalence of online interviews. *"This came at a really good time.... trying to get my head around the technology, and now it's all video [online] interviews rather than face to face."* [F, peer, 37]. This persuaded participants to acclimatise to the technology, encouraging older participants to display initiative. *"What I think I'm going to do is: I'm going to learn from my nephew how to tape yourself on the computer"* [M, peer, 61]. Several mentioned that they had **experience with AVIs as a part of their job applications**. These external factors can influence their positive usage of InterviewApp. Moreover, external factors like **imminent interviews motivated them to train** further. An upcoming interview or the anticipation of one persuaded them to record and review videos.

On the other hand, the **lack of upcoming interviews**, obtaining a new job, and personal priorities limited usage and intent to train. *"I think I would have practised if I had an interview coming up."* [F, self, 37]. Several participants also mentioned other factors, such as **interviewer behaviour, job role interest**, and the relative competition of other candidates influencing the performance. These factors affect the performance of the participant during the interview, irrespective of their training efforts. *"The person that you have in front of you, sometimes you have a more friendly person that makes you feel at ease. So that definitely plays a big role in an interview."* [F, self, 35].

Most participants recognized the **impact of their emotional state** on performance, including nervousness and anxiety during interviews. The pressure to excel in the limited interview callbacks, despite applying for many jobs, could exacerbate performance anxiety. *"Lots of interviews, lots of applications! The vast majority of applications ended with no response whatsoever, and it becomes very disheartening."* [F, expert, 37]. Similarly, the effect of emotional state was apparent to the participants during training. Successful self-review and self-evaluation require a conducive mindset during training sessions. *"You're having to judge yourself again. 'Am I feeling really confident today? Do I think I did good, or do I feel like kind of meh today? Did I just go meek?' "* [F, expert, 26]. Overall, various external factors might have attenuated the effect of our intervention, leading to restricted autonomy and a lack of influence on performance.

2.6.5.2. Theme 2: Autonomous satisfaction This theme provides insights into the satisfaction of autonomy. It encompasses the: *(i)* need for autonomy in job search and training,

(ii) autonomous satisfaction through InterviewApp, and *(iii)* need for more autonomy-supportive features in InterviewApp.

Several participants emphasised **owning their training process**. *"I would create an area where a person would be able to talk about what they're passionate about. And how that relates to the job that they might be applying for."* [M, self, 43]. Most engage in **self-directed training activities** such as researching the company and job profile, rehearsing questions, and watching instructional videos. A few participants even mentioned that they devise unique methods; an artist explained how their approach differs from typical preparation. *"It's so different with what I do and what I usually have to prepare for, that I find my own ways to do it."* [M, self, 43]. Participants expressed a desire to exert control over their job search. Several mentioned that they are **choosy over the job roles**. *"So I think I was definitely being fairly choosy. Over the last four months, I've only had two interviews."* [F, self, 35].

InterviewApp satisfied this desire for autonomy, to an extent. About half of the participants we interviewed **preferred online solutions** like InterviewApp for its **convenience and flexibility**, recognizing its potential for self-directed learning opportunities. *"A video service [like InterviewApp] is a little more convenient rather than scheduling a meeting with someone separately [to get feedback]."* [F, peer, 27]. Several participants found InterviewApp's **consistent content across features** and the training guide to be valuable, serving as compelling reasons for their interest and engagement in the training sessions. *"I was already aware of suggestions [in the guide], still a good way to break to you what you have to focus on before doing the interview."* [F, peer, 35]. Additionally, InterviewApp's versatile content, including questions **adaptable to various profiles**, was seen as a significant advantage. *"Because the questions were flexible, I always pretended interviewing as a videographer."* [M, expert, 23].

Several participants expressed the need for greater autonomy through **more choices, improved explanations, and increased transparency**. They requested additional choices like video deletion, job profile selection, and adding their own questions. *"If there were the option to record a mock interview and then choose whether you could discard it would be good."* [F, self, 29]. While most participants appreciated personalized comments on the video, about half desired greater transparency concerning the feedback generation process. *"You may prefer the interviewer [expert] telling you actually how he arrived at those parameters."* [M, expert, 38]. About half of them called for **more personalized training experiences**, including tailored questions based on performance and job profile. *"It will be nice if you could select what kind of job role going forward, to get more specific things [questions] come up"* [F, self, 44].

2.6.5.3. Theme 3: Criteria for competence In this theme, the participants discuss two key aspects. Firstly, they elaborate on how InterviewApp strengthened their sense of competence through feedback. Secondly, they provide suggestions to enhance the feedback process and ultimately improve their competence.

Almost all the participants acknowledged that the feedback received through InterviewApp was actionable and pertained to their current performance. *"InterviewApp got me to think about some aspects of interviews and what to say and what not to say. So I feel*

that, with that, I was able to perform a little better." [M, expert, 38]. Several mentioned that **overall comments is the most significant takeaway**. Additionally, several noted that the **feedback was very detailed**. *"I didn't expect it to have gone into so much detail"* [M, 23]. About half of the participants reported receiving specific areas of improvement in the feedback, which they could **incorporate into future interviews**, finding it **valuable for real-world interviews**. *"In my interview for the university, I was really proud of the way I handled myself and part of that [was because of] practice I had done with the InterviewApp a few times and gotten feedback."* [F, expert, 26].

Furthermore, about half of the participants expressed their desire to **prepare well in advance** for interviews, with several mentioning that InterviewApp played a crucial role in this preparation. *"If you're practising for an interview, generally speaking, I'm practising the week of the week before... And then the day of the interview, I don't prepare at all"* [F, self, 29]. However, several emphasized the value of spontaneity during interviews, cautioning against over-preparing. *"If you're not necessarily super prepared for something you get a bit better [...], it just feels like you're a bit more yourself"* [F, peer, 27]. A few also expressed their aversion towards receiving negative feedback immediately before interviews, believing it to be potentially counterproductive.

Nearly half of the participants made additional recommendations to enhance the feedback, potentially augmenting their competency and performance. Several participants expressed the desire to **progress from a lower score to a higher score** for each job attribute, seeking targeted feedback. *"It probably would be nice to have links, maybe to some information about how to improve the specific field [attribute]."* [F, self, 35]. Participants indicated an interest in **feedback that includes examples** showcasing ideal or sub-optimal interviews. *"If there were like a little short video where somebody has done that, and it would spark some ideas on how you could incorporate that a bit more into your technique."* [F, self, 44]. Furthermore, **cumulative feedback showing progress over time** was deemed beneficial by a few participants for tracking performance across the mock sessions. The desire for more comprehensive feedback incorporating these additional features might have mitigated the direct impact of the experimental manipulation on competence and performance.

2.6.5.4. Theme 4: Need for relatedness This theme reinforces the findings detailed in Sec. 2.6.1.1 and Sec. 2.6.1.2. It mainly comprises the following aspects: *(i)* impediment in peer support, *(ii)* acknowledgement and support through feedback in InterviewApp, and *(iii)* desire for human-like elements to compensate for the lack of human contact.

Most peer group participants indicated a **lack of supportive peer network** and hence could not request feedback. *"The main reason I don't go to other individuals when practising is the jobs I apply to are very niche."* [F, expert, 26]. Several participants felt that the **peer feedback quality is lesser than an expert** as detailed feedback could not be expected from peers, given the informal circumstances. *"Feedback from a friend or a colleague might not be as relevant as opposed to someone if I'm getting it from a professional perspective."* [M, expert, 37]. A few participants also felt uncomfortable practising with acquaintances. A few recommended creating a **pool of peers to receive feedback** in training. *"People that are looking for interviews could give each other feedback [...] So we can all be constructive and help each other."* [F, peer, 37]. Prior work has shown some success in this regard

[25, 320]. (See more in Sec. 2.7.2).

About half of the participants **preferred expert feedback** and **appreciated the anonymity**. *"I wasn't sitting there thinking about how someone was going to judge me [...] it felt genuinely constructive."*. [M, expert, 37]. Several peer and expert group participants favoured feedback from unbiased sources, with whom they were not personally familiar. Most of the participants who received feedback reported **feeling acknowledged and supported by the feedback** they received. *"It was actually interesting to know that they actually paid attention to certain areas of the video they watched."* [M, expert, 38]. Further, a few participants sought additional support, such as expert input via video, multiple expert opinions, or a combination of expert feedback and peer feedback.

However, most participants acknowledged the **absence of human interaction** and social reciprocity in online and video interviews. *"You don't get that [rapport] with a video interview, and that's what's really difficult because you're just looking at yourself."* [F, expert, 37]. They noticed a lack of prompting and turn-taking. They proposed various strategies for adding a human touch to compensate for their need for relatedness, such as **real-time hints or prompts** and/or using avatars instead of no person on the other side. *"If you haven't answered it [a question], they can kind of prompt you."* [F, peer, 37]. There have been attempts to enhance the social interactions in video interviews by having avatars ask the interview questions [14, 143] or by prompting follow-up questions to nudge respondents to provide more complete answers [254].

2.6.5.5. Theme 5: Self-reflection - a bitter pill to swallow This theme provides insight into participant perceptions of self-reflection, supporting our findings in Sec. 2.6.1.3 and Sec. 2.6.1.2. About half of the participants agreed that self-review fosters self-awareness. First, a dimension of self-reflection was added by the **self-view in the camera** while the participants recorded the mock interview. *"How would you look, what you are saying, what are the things that come across well. I can actually read into it a lot myself"* [M, peer, 61]. They mentioned that looking at yourself on camera is uncanny at first, but gets better with practice. Additionally, the self-view helped with eye contact. *"It's good, it helps with where to focus my eyes"* [M, self, 43]. Second, several participants noted that **self-reflection through self-evaluation** helps.

Nonetheless, participants did not frequently use the playback feature, as most **disliked watching themselves in videos** and found it to be an uncomfortable experience. *"Do the practice ... and then left it at that. I definitely didn't go back and watch any of them because, if anything, that would make me feel far worse."* [F, self, 29]. Several other participants also felt that self-reflection through self-evaluation was not a positive experience. Although they **recognized the benefits of self-review**, about half of the participants were unable to practice it due to the perceived embarrassment associated with the experience. *"Probably it would be helpful for me, but I would find it really stressful. It's just really embarrassing."* [F, self, 37].

2.7 Discussion

This study demonstrated the effect of training through InterviewApp on the basic psychological needs for interview motivation and, in turn, on performance. We begin by discussing the salient results of our study, which highlight the importance of interviewee motivation in influencing performance and its implications for design. Our findings indicate that the InterviewApp intervention had a significant effect on the job seekers' autonomy. The request feedback feature supported autonomy and relatedness. Further, we found that autonomy is a mediator of performance. The engagement with InterviewApp using recording and feedback features had a positive impact on performance. The feedback page views and the detail views had a positive impact on communication, composure, self-promotion, and confidence. However, InterviewApp was not found to be effective in fostering peer support.

2.7.1 Autonomy as a Mediator of Performance

Prior literature [96, 314] has established the significant role of autonomy and self-regulation in facilitating job search. The presented findings demonstrate that *autonomy mediates the effect of training on interview performance*. Training for job interviews was found to influence job interview performance, and the size of this effect could be explained by the autonomy possessed by the job seeker. We found this relation between autonomy and performance to be *negative*. Note that participants with higher levels of autonomy (or who achieved higher levels of autonomy during the experiment) were more likely to be *intrinsically* motivated towards finding a new job, while those with lower levels of autonomy (or participants whose autonomy decreased during the experiment) were more likely to be *extrinsically* motivated towards finding a new job. Individuals who are intrinsically motivated toward finding a new job are those who might think about long-term job perspectives. In other words, those might be driven toward finding a job that better resonates with their passions (i.e., their dream job) and towards the betterment of their life. Individuals who are extrinsically motivated towards finding a new job are those who might be, given their life circumstances, more in the immediate needs of finding employment, whatever that might be, or those who suffer more from the societal pressures of being unemployed and wish to exit unemployment as soon as possible. Previous studies have contended that these *extrinsic motivators* might be needed to maintain involvement in the task activity and sustain performance until intrinsic interest towards the activity develops or a certain level of proficiency is achieved [108, 182]. This study shows that job seekers engage in the activity of training for their next job interviews with varying efforts and that the effort they manifest is related to the level of autonomy they perceive to possess towards finding a new job. The presented results show that extrinsically motivated participants devoted more effort to training (more logins, longer duration, and more feedback views) through InterviewApp than intrinsically motivated participants and achieved higher results in some of the performance metrics than participants who possessed higher levels of autonomy. These job seekers spent their efforts on training to satisfy external demands to achieve their short-term objectives of securing a job. While extrinsically motivated participants focused efforts on training for their next job interview, intrinsically motivated participants might have devoted cognitive resources towards other activities (e.g.,

finding the most suitable job openings that matched their interests and needs).

For some job seekers, extrinsic motivators could provide sufficient motivation to secure a job. However, recent studies have shown that maintaining a high level of autonomy during the unemployment spell is crucial to accrue *transferable skills*, which might span onto the next job searches and foster *continuous growth* of the job seeker [59]. Autonomy encourages independence and decision-making, which may make it harder for individuals to appreciate its immediate positive effects, especially in situations with delayed results (e.g., the benefits of interview training may not be evident until securing a job). However, with time, as job seekers engage in self-directed learning and tailor the training content to their specific job requirements, they recognize the benefits of autonomy-based training [59]. When job seekers aspire to acquire transferable skills for enhanced job performance and pursue their *dream job* beyond mere unemployment exit, training autonomously becomes imperative. It is important to note that job seekers recognize the importance of autonomy in their training process: all job seekers we interviewed desired more autonomy in their training process.

The experimental manipulations had an effect on the levels of autonomy perceived by the participants. Participants who had the possibility to request feedback reported higher levels of perceived autonomy than the other participants at the end of the experiment. However, it must be noted that –overall, the effect of experimental manipulations on autonomy might have been crowded out by the effects of other exogenous variables that we did not control (e.g., critical life circumstances, loss of subsidies, the income of other household members), not to mention the constraints posed by the experiment on participants' lives. We will come back to this point in the limitations section (see Sec. 2.7.4).

The main implication of our research is that autonomy plays a key mediating role in the design of online interview training tools to influence the motivation of job seekers and interview performance. Future designs of interview training tools must consider incorporating this novel insight by adding extra features that enhance autonomy. For instance, this could be achieved by providing greater transparency, customization, and personalization. Some suggestions that can encourage choices and nourish volitional action include: *(i)* enhanced transparency regarding feedback generation, for instance, anchoring the feedback to specific points in the video and explaining a certain score; *(ii)* more flexibility with video recording, for example, allowing preparation time before starting the recording, ability to re-record or delete videos; *(iii)* an option to tailor interview content, for example, questions can be based on the job seeker's personal profile.

On a broader level, this study highlights the importance of designing *support for longitudinal life events* that could be separated by multiple years (e.g., distinct unemployment spells). Most technological supports for job seekers typically focus on isolated occurrences. However, future designs might consider how job seekers could transfer knowledge accrued in one event onto the next event. Future tools for employment should support job seekers directly (i.e., by improving interview skills, addressing immediate needs) and indirectly (i.e., by inspiring autonomy on their career path and exploration, catering to long-term needs). Designs of such tools could have a broader scope focusing on transferable skills moulding into unique contexts (like pivoting careers, changing job roles, getting back to education, etc.) and remain relevant over time. Some examples include: *(i)* skills such as communication and storytelling, valuable in interviews, also play essential roles on the job; *(ii)* practicing

responses to common interview questions, such as 'introduce yourself', can also aid in delivering an effective elevator pitch or networking effectively; *(iii)* history of mock interview videos can help in comparison of performance across different unemployment spells. Given the dynamic job market influenced by technologies like Artificial Intelligence, employment tools should consider empowering job seekers' autonomy to tailor content for their current and future jobs. Furthermore, job seekers should recognise that autonomy's greatest impact manifests during on-the-job problem-solving, emphasizing the need for a long-term approach to assessing training outcomes to comprehensively capture the varying impacts of autonomy.

2.7.2 Lack of Peer Network

Another important finding of this research is that the peer condition was unable to provide support to the relatedness needs of the participants. These findings resonate with earlier research on digital employment tools [94, 96] where it has also been found that peer networking features did not lead to useful support. On the other hand, the request feedback feature had a notable impact on the sense of relatedness among job seekers, as participants in the expert group could greatly benefit from it. However, the peer group encountered challenges in either successfully requesting feedback or receiving it when requested. The primary causes were that the job seekers lacked a supportive peer network or refrained from requesting feedback, as they believed peer feedback would not be as insightful as that of an expert. Several interviewed participants reported preferring an evaluation from a stranger, which they considered *unbiased*.

Receiving non-actionable feedback has been found to have a detrimental effect on the relatedness need of the receivers [289]. Feedback from a colleague or friend can be a set of unstructured suggestions. Therefore, we designed feedback that job seekers could potentially receive from peers to be structured and actionable on a number of job interview criteria, hence supporting their relatedness. To help them understand the significance of this structured feedback, such as communication and composure, we provided examples of feedback they could receive. However, the job seekers either missed these examples or believed that peer assessment of this nature would be highly subjective. Not receiving any feedback (despite an issued request) might have a detrimental effect on the relatedness need of the person requesting the feedback. In addition, this may have added to the frustration of those job seekers who could be already dissatisfied with the lack of responses to their real job interview applications.

According to prior research, creating chances for job seekers to learn from and support one another is essential in establishing a positive learning environment [320]. Therefore, we learn from our experiment that forcing job seekers to solicit input from peers *already* in their own network does not provide support for their relatedness need. However, helping them to connect with *other* peers, who might perhaps be going through the same experience, could build a sense of trust and community. Additionally, pairing job seekers with other job seekers might accomplish both goals of getting an unbiased evaluation from an outsider and creating a welcoming environment that encourages relatedness.

Going forward, we advise designers to include capabilities that could promptly link job seekers with people who might offer advice or connect them to similarly experienced job

seekers to build a collaborative and supportive environment. Matching them to other job seekers using the tool based on their profile or industry domain might increase the relevance of the feedback and help bootstrap the community. Prior work showed that these cooperative interactions can scaffold learning and increase self-reflection [25, 97]. Further, tools can be designed to assist in interpreting diverse peer feedback received by filtering out low-quality feedback, detecting patterns, identifying contradictions, and subsequently adding a coherent structure to the received feedback [344].

2.7.3 Influencing Interviewee Performance

The design of InterviewApp positively affected job seekers' performance. Job seekers who practiced by recording mock interviews, reviewing the feedback, and implementing the suggested changes improved their job interview performance, as measured at the end of the experiment. Participants in the experimental conditions who recorded longer mock interviews managed to improve various aspects of their answers, such as weaving stories, maintaining composure, and communicating clearly [147, 151]. Spending more time providing complete answers was found to have a greater impact on their final outcome than practising repeatedly on different batches of questions.

The playback feature positively influenced the composure and confidence of job seekers. These two attributes, especially connected to the interviewee's appearance, pertain to the applicant being relaxed, at ease, calm, and comfortable [142, 147]. Participants in the experimental conditions who practised through InterviewApp could potentially self-correct their appearance with the use of the playback feature. Some previous studies have indicated that interviewees do not always prefer self-monitoring via videos [144, 269]. It is interesting to note that the playback feature, which was also not liked by many participants in our study, was indeed able to provide useful support to job seekers during their practices.

InterviewApp also helped prepare for online interviews. Interviewed job seekers reported that it was more challenging to convey their personalities in online job interviews than in in-person interviews. For this, the video record feature with self-view in the camera during the practice sessions with InterviewApp proved useful.

Furthermore, this study also sheds light on the design of the feedback that video interviewing sites should provide to job seekers. We discovered that the feedback was useful in attracting the trainees' attention to certain skills (job attributes), which are typically looked at by recruiters in job interviews. The feedback detail view helped job seekers assess their performance with regard to these skills and helped them determine how to best improve on these specific aspects of their interview. The overall comments were reported to be mostly actionable. Overall, the feedback was successful in enhancing performance (fulfilling immediate job-seeking needs). However, the feedback was not able to increase the participants' perceived level of competence. The job seekers aimed to attain mastery in specific job attributes and sought more comprehensive feedback, with concrete examples, to understand how to improve their scores on each attribute.

We recommend future designs of online job interview training sites to provide informational and actionable feedback for each attribute, with examples to satisfy the competence needs of job seekers. Some ways of providing actionable feedback with examples include

video examples of an ideal interview, behavioural examples to illustrate how the attribute can be demonstrated during an interview, tutorials for mastering a certain skill, and personalized feedback based on job seekers' strengths and weaknesses. As also discussed in Sec 2.7.1, future designs should work toward increasing the informational value of the feedback by enhancing the transparency and explainability of the feedback [255]. Some examples include video feedback explaining the scores and highlighting specific elements in the answers that led to the score.

It must be noted that the way we generated feedback in this experiment does not scale, and it is cost-inefficient due to the reliance on the availability of experts (i.e., job interview trainers or professional recruiters). Limited counsellors' availability and scheduling conflicts can hinder access to timely feedback. In this regard, automatic assessment has the potential to narrow this gap [135, 219, 224, 256, 261]. As many of the job seekers we interviewed preferred expert assessment, we advise combining automated feedback with human evaluation. Initial training can be assisted by automated evaluations to help job seekers get started, and when their scores reach a certain proficiency level, a final evaluation with professional assistance can be made to provide definitive recommendations to the job seekers.

2.7.4 Limitations

Our study has some limitations. Any experiment involving human subjects might have constraining effects on its participants, impacting their perceived level of autonomy. To limit this effect, we informed participants they could access InterviewApp as much or as little as they wanted, as this was not going to impact the incentive they were going to receive at the end of the study. However, participants were required to complete specific activities required by the experiment (e.g., baseline and post-treatment mock interviews, diary entries), and their ability to choose to exercise autonomy in various aspects might have been curtailed. While these activities might have contributed to a reduced sense of autonomy perceived by the participants, they equally impacted –by design– all experimental conditions. It must be noted that being unemployed poses a far greater load on the perceived sense of autonomy, which might have crowded out the smaller effects discussed in this paper.

Our study was tied to the Prolific recruitment platform. All communication occurred via Prolific's messaging system. Direct contact with participants was restricted, preventing us from ascertaining why some participants stopped participating in the study; perhaps they found a job, while others simply might have lost their Prolific password. For instance, a participant who rehearsed with 13 mock videos did not complete the post-treatment session. To better evaluate attrition related to the use of online video interviewing platforms, future research could recruit job seekers through channels that allow direct communication. For instance, these types of participants can also be recruited through regional employment offices.

On average, job seekers in the study appeared for approximately two real-world job interviews in the entire study duration. A four-month study period could be insufficient for job seekers to receive responses to their job applications, leading to a lower number of interviews. The uneven demand in the labour market during the study period [234] could have exacerbated this issue, potentially resulting in diminished need to prepare for job interviews and

consequently explaining the observed low levels of engagement with the InterviewApp (lower number of logins, mock interviews, and feedback requests). We recommend that researchers interested in pursuing similar research design experiments with a duration of six months or longer, during different labour market conditions in the post-pandemic era.

In our analysis, while several regression models exhibited statistically significant results in predicting the outcome variables, certain models displayed low adjusted R^2 values. Such findings suggest that the variables accounted for explain a small portion of the variation within the model, indicating the presence of unobserved factors contributing to the residual variance. This is further supported by the qualitative findings (Sec. 2.6.5.1). In future research, exploring additional variables beyond those considered in the current analysis may enhance the predictive power of regression models.

Our study relies on a single expert to rate all pre-treatment, post-treatment, and feedback-requested videos. We took measures to enhance the reliability and consistency of our subjective assessments by implementing a rigorous training and calibration process involving multiple raters. The high ICC values for a subset of videos affirm the robustness of our assessments. Nonetheless, we acknowledge that involving multiple raters in future research could further enhance the generalizability and objectivity of our findings.

One of the job seekers who participated in the qualitative interviews had a diagnosis of autism. This participant mentioned that the feedback from InterviewApp could come across as quite ablist. This comment made us realize that our design did not undergo a formal accessibility evaluation. Future research and design might want to design ways to make online interview training platforms more inclusive and perhaps study how to best train differently-abled job seekers.

2.8 Conclusion

In this paper, we introduced InterviewApp, an online video interview training tool designed to support job seekers' motivation and performance. Through a large-scale deployment of InterviewApp, we demonstrated the mediating role of autonomy on the effort job seekers manifest to train for their interviews and the related impact on their interview performance. Through the study, we learned how the design of video interview training platforms, such as InterviewApp, caters to the immediate goals of job seekers and is able to support their short-term goals of boosting performance. Furthermore, the presented findings shed light on the complexities of designing interventions that simultaneously empower job seekers' sense of autonomy and help improve their performance. Maintaining a high level of autonomy while training could help job seekers address their long-term goals and needs.

Future research should focus on developing and testing alternative designs for online video training platforms that could support short-term needs while—at the same time—increasing job seekers' autonomy and helping them gain transferable skills. We plan to move in this direction through a participatory design study, involving job seekers in the design of a training platform that could support multiple types of employment events. This study could also help gain insights into their specific requirements for different job attributes and help improve the design of feedback. Furthermore, we also plan to design enhanced support for the relatedness

of job seekers. Specifically, the peer evaluation feature could be redesigned to leverage the community of users of the online video interview platform rather than requiring job seekers to identify peers in their own network.

As companies are rapidly changing strategies and methods to recruit human resources, more tools are needed to support job seekers. We hope that the presented study will stimulate the design and research of employment technology that could empower job seekers, protect them from technological discrimination, and help them go through these life-changing events with more ease and satisfaction.

Chapter 3

Involving Job Seekers and AI Practitioners in Co-designing AI-assisted Interview Training Systems

Abstract. Recent advancements in AI have expanded its use in training systems, including job interview training. Despite the call for participatory approaches in AI design, many interview training systems often overlook stakeholders' involvement, missing varying needs and perspectives. To address this, we employed a multi-phase co-design methodology engaging two key stakeholder groups. First, *N = 18* AI practitioners in a focus group provided insights on AI applications in interview training. These insights were presented to *N = 20* job seekers in co-design activities, resulting in several tailored designs. Finally, *N = 10* AI practitioners evaluated these designs and provided feedback on feasibility and improvements. Our findings revealed that both stakeholders aligned on preferences like actionable feedback and transparency, while differing slightly on accountability and human-AI collaboration. Results also highlighted four socio-psychological risks perceived by job seekers: homogenization of communication, overreliance on AI, illusion of human interaction, and reinforcement of stress. These findings offer valuable design considerations for AI training systems, demonstrating the significance of stakeholder engagement for effective user-centered solutions.

3.1 Introduction

Rising unemployment, exacerbated by global economic challenges, remains a critical issue. According to a report by the International Labor Organization [152], globally, more than 183 million people are unemployed in 2024. Job interviews remain the most common selection method for employment [307], and digital platforms are being increasingly used for hiring [109]. The rapid proliferation of Artificial Intelligence (AI) technologies, combined with

the declining costs of their adoption, has led to their increasing application, fundamentally changing the workplace and recruitment [235].

Online interviews for hiring are conducted through computer-mediated communication methods synchronously, near-synchronously, or asynchronously [275]. Asynchronous video interviews (AVI) or online video interviews [174], enable candidates to record responses to textual prompts, allowing them to complete interviews at their own convenience. These interviews are gaining popularity due to their scalability and ease of use. AVIs are used in early screening rounds[26] and provide an opportunity for interview training, leveraging AI techniques like computer vision, natural language processing, and audio processing to provide automated feedback. These systems are typically multimodal, utilizing data from various sources and modalities [219, 224, 256]. Beyond AVIs, this broader category of AI-assisted interview systems also includes synchronous formats such as virtual avatars and chatbot-based interviewers, which aim to create more interactive and immersive environments [14, 77, 303].

AI-assisted interview systems can be leveraged as self-help tools, reducing the dependency on the availability of trained counsellors for interview training [257]. Several studies in the literature employ them for automated training and feedback, not only for job interviews [135, 173] but also for developing other soft skills like communication [243], social skills [143], and public speaking [115]. While these AI-assisted training systems offer the potential to enhance job preparedness through feedback and interactive training, their design often overlooks user needs.

The significance of involving job seekers in designing employment tools has been well-established [92, 93, 95, 136]. The use of AI in automated interview training introduces novel and complex needs for job seekers, yet very few studies have explored these [159, 175], particularly during the design phase. There is a growing call for participatory approaches in AI system design and development to involve members of communities and stakeholders impacted by the AI systems [84, 342, 345]. Yet, existing surveys show that stakeholder involvement in AI design is often shallow, largely consultative rather than *collaborative*, and rarely shifts decision-making power [84, 112].

Such limited involvement is particularly problematic in high-risk domains like hiring, where a *power dichotomy* exists between developers and deployers of AI systems (referred to as AI practitioners) and the people directly affected by the systems (job seekers) [198]. Hiring and training systems affect the livelihoods of job seekers, making it essential to redistribute power in design processes and elevate their perspectives. Bringing job seekers and practitioners into the same design process can surface tensions between technical feasibility and user desirability, flag potential concerns, and enable joint shaping of system goals, features, and safeguard mechanisms.

Further, designing human–AI interactions poses significant challenges. Difficulties in understanding AI can prevent job seekers from understanding its potential benefits and may lead to either unrealistic or diminished expectations [166, 185, 345]. Technical expertise from practitioners is therefore crucial to ground design ideas in what is feasible and reliable, while also ensuring responsible practices around transparency, fairness, and risk mitigation [200, 324]. Participatory processes that integrate both groups can thus uncover not only

[26]For example, `www.hirevue.com`, `www.usebraintrust.com`, `www.viasto.de` Last accessed Sept 2025.

functionality requirements but also ethical and social considerations. Despite these calls, the existing literature on automated interview training [14, 135, 143, 243, 256] largely ignores stakeholder engagement in design.

To address these challenges, we conducted a multi-phase co-design study, actively engaging job seekers and AI practitioners in designing AI-assisted interview training systems. The goal of the co-design study was to understand job seekers' needs and concerns, while also drawing on practitioner expertise to anticipate responsible AI and feasibility challenges. Through this exploratory study, we specifically aimed to answer the following research questions :

RQ1. What are the needs and concerns of job seekers regarding AI-assisted interview training?

RQ2. How can insights from AI practitioners inform the feasibility and responsible design of such systems?

We conducted the study in three phases, each building on the outcomes of the previous: a *focus group* with 18 AI practitioners to explore applications and pitfalls of AI in interview training (Section 3.3.1); a *co-design workshop* with 20 job seekers to elicit needs, concerns, and design ideas (Section 3.3.2); and an *online evaluation* with 10 practitioners to assess the feasibility of job seekers' designs and gather responsible practices (Section 3.3.3).

Our findings highlight a broad set of design requirements for AI-assisted interview training tools. Both job seekers and AI practitioners emphasized the importance of systems that adapt to individual needs (personalization), embed human-like design features (anthropomorphism), provide clear and actionable feedback, work with diverse data, and remain transparent about how they operate. While both stakeholder groups broadly aligned on these priorities, a few notable differences emerged around accountability, human-AI collaboration, and long-term risks. Job seekers, for example, stressed the need for human oversight, the ability to challenge or correct AI-generated outputs, and safeguards against risks, concerns that practitioners began to recognize only after integrating user perspectives. Job seekers also perceived four socio-psychological risks of AI-assisted training: **homogenization of communication, overreliance on AI, illusion of human interaction, and reinforcement of stress**. These findings contribute to empirical insights into design requirements and the user-perceived risks associated with AI-assisted interview training. In doing so, we offer considerations of broader applicability for human-AI interactions in training tools.

3.2 Related Work

In this section, we will first review works on AI in interview training and feedback, followed by the literature related to designing employment tools.

3.2.1 AI in Interviews and Job Training

Rapid advances in AI have led to the automation of recruitment and interview training tools. Langer et al. [173] introduced and evaluated virtual interview training in comparison

to traditional methods like sample questions and video presentations. Similarly, Muralidhar et al. [218] designed a behavioural training procedure for hospitality students to enhance their social skills in job interview scenarios, focusing on first impressions. These studies suggest that practising interviews with virtual avatars improve skills and performance more effectively than traditional methods such as books or videos [77, 173].

Advances in social signal processing [319], involving the analysis and synthesis of social signals in human-machine interactions, have further enhanced AI-based interview training systems by incorporating multiple modalities such as audio, video, and text [135, 219, 224, 256, 261]. Some tools use virtual avatars to provide feedback on nonverbal behaviours, such as body language and facial expressions, during or after simulations [14, 143]. For instance, `MACH` [143] is a virtual social skills coach in job interview training, while `ASST` [303] focuses on social skills development. Virtual avatars create a more interactive and immersive training experience, adapting their behaviour to simulate a real interviewer with features like follow-up questioning [254], social attitudes [54], and stress induction [118]. Studies also show that virtual interviewers increase self-disclosure and reduce evaluation anxiety in job seekers [193].

While these studies have developed and assessed automated interview training tools, they did not involve end users in the design and development stages. Involving stakeholders is crucial to understanding their needs and improving these tools' effectiveness, which is the focus of our study.

3.2.2 Stakeholder Involvement in Designing Employment Tools

The importance of involving stakeholders in the design of employment tools is well-established. Dillahunt et al. [92] investigated the needs of disadvantaged job seekers and insights from recruiters to design a tool for resume feedback, helping users reflect on their skills. Building on this, Dillahunt et al. [95] provided design insights for tools supporting feedback, interviewing, and related activities, validating concepts tailored to underserved populations. Similarly, Dillahunt et al. [95] conducted a needs validation study to assess the viability of ten different design concepts drawn from literature. They found that tools for reviewing feedback, identifying skills were most desired and provided design insights for employment tools tailored to underserved populations.

Hendry et al. [136] highlighted the need for emotional support and external resources, showing how user-centred designs can address both personal and societal challenges. Dillahunt et al. [93] noted how rapid technological changes disproportionately affect marginalized job seekers, underscoring the role of employment centre staff and proposing a framework to design tools responsive to these unique challenges. Other studies [94, 257] apply psychological frameworks like Self-Determination Theory [271] to design and evaluate how training tools can support autonomy and self-reflection.

A smaller body of research addresses AI in employment tools. Lashkari and Cheng [175] compared recruiter and job seeker perspectives on automated recruitment tools, emphasizing the need to mitigate cognitive biases and support job seeker agency. Charleer et al. [64] designed and evaluated an interactive dashboard through a user-centred approach to help job mediators explain job recommendations, thereby improving communication and decision-making.

3.2.3 Participatory and Co-Design Approaches in AI Systems

A growing movement in AI system development advocates for approaches that go beyond traditional user-centered design to more deeply engage stakeholders in shaping the direction and purpose of these systems: a shift often referred to as the *participatory turn in AI design* [84]. The shift from user-centered to stakeholder-centered design recognizes that AI systems operate within complex socio-technical ecosystems where multiple actors interact [112, 180, 230].

Delgado et al. [84] analyzed 80 participatory AI projects and interviewed researchers with experience in participatory methods. They found that only 10% of projects empowered stakeholders to influence system goals or define model features. Most projects relied on one-off input methods such as interviews, surveys, or workshops, and only a handful [121, 178] adopted a multi-phase engagement to meaningfully involve stakeholders across the lifecycle.

Recent efforts have begun to shift towards more inclusive co-design practices. WeBuildAI [180] is a participatory framework that enables community members, organizational staff, and technical facilitators to build algorithmic policy. The approach operationalizes multi-party value deliberation and shows improvements in perceived procedural fairness and outcomes. Bogucka et al. [38] engaged AI practitioners and compliance experts in an iterative co-design process to develop an AI Impact Assessment Report Template. This multi-stakeholder co-design led to a template grounded in regulatory frameworks, which was later evaluated positively by additional practitioners and experts. Kawakami et al. [160] worked with a multi-disciplinary group including AI developers, public sector agency leaders, frontline workers, and community advocates to co-design the Situate AI Guidebook. Through an iterative co-design process, they created a toolkit to support early-stage decision-making for public-sector AI projects.

Research Gap. Prior research demonstrates the importance of involving job seekers in the design of digital employment tools. With the growing use of AI, such involvement becomes even more critical for uncovering the nuanced needs and concerns of job seekers. Engaging both job seekers and practitioners in the same design process is crucial, as it can expose tensions between technical feasibility and user expectations, redistribute decision-making power, and foreground issues of fairness and accountability. However, current work on AI-assisted interview training has largely overlooked such participatory approaches, leaving open the opportunity to explore how stakeholder involvement can inform the design of effective and equitable AI-assisted interview training systems.

3.3 Method

We employed a multi-phase co-design methodology to investigate the requirements and challenges faced by job seekers using automated interview training systems and analyse their alignment with expert inputs on AI capabilities and feasibility. Our study consisted of three phases (Figure 3.1): 1. a *focus group* to gather insights from AI practitioners on the application of AI models for automated interview training (§3.3.1), 2. a *co-design workshop* to identify requirements and understand job seeker needs for such a system (§3.3.2) and

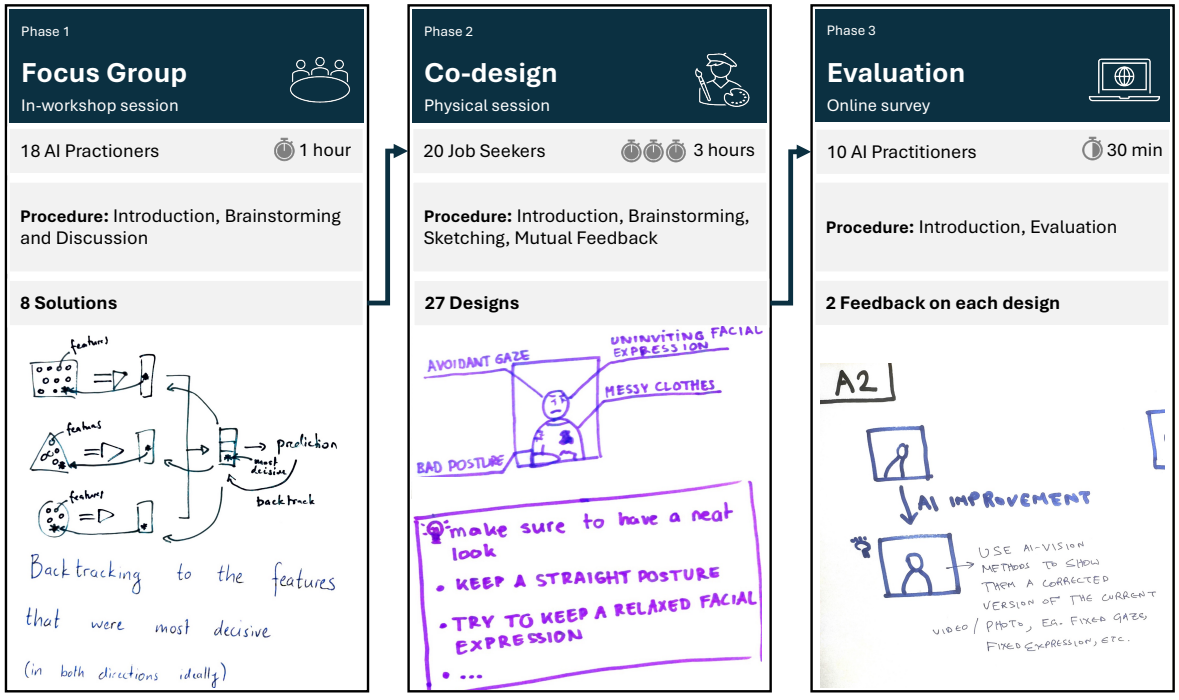


Figure 3.1: Our study was conducted in three phases: (1) A *focus group* with 18 AI practitioners to gather insights on the application of AI models for automated interview training. (2) A *co-design workshop* with 20 job seekers to identify requirements and understand their needs for an automated interview training platform. (3) An *online evaluation* with 10 AI practitioners to gather feedback on the job seekers' designs. The solutions from Phase 1 were combined and presented as concepts for an initial proof-of-concept approach in Phase 2. Each design from Phase 2 received feedback from two AI practitioners in Phase 3. The illustrations within each phase show examples of updates made throughout the process. For instance, design A2 in phase 2 evolved from textual ***AI feedback*** to ***AI improvement*** using an AI-generated ideal version of the current response.

3. an *online evaluation* by AI practitioners to gather expert feedback on the job seekers' designs (§3.3.3). Each phase in the study was informed by the outcomes of the previous one. We began with insights from AI practitioners in Phase 1 to ensure that job seekers' design activities would focus on features that are feasible within current AI capabilities.

3.3.1 Focus Group Workshop with AI Practitioners

We organised a multimodal machine learning seminar at a conference attended by industry professionals and AI practitioners with expertise in AI and machine learning (ML). Within the framework of this seminar, a hands-on demonstration was conducted, showcasing the development of an AI model designed to evaluate job interview performance in AVIs. During the final hour, we conducted a focus group (phase 1) to engage AI practitioners and gather insights on the development and application of ML models in automated interview training and feedback systems.

Table 3.1: Self-reported knowledge and demographic characteristics of the participants across the three phases. AI use was rated on a 1–7 scale (1 = never, 7 = very frequently), and we report the mean (±SD) usage. Practitioners reported AI use at work, and job seekers reported AI use in everyday tasks.

	Characteristic	**AI Practitioners (N=18) Phase 1**	**Job Seekers (N=20) Phase 2**	**AI Practitioners (N=10) Phase 3**
Education	Bachelors	-	50% (n=10)	30% (n=3)
	Masters	72.2% (n=13)	25% (n=5)	60% (n=6)
	PhD	27.8% (n=5)	5% (n=1)	10% (n=1)
	Other	-	20% (n=4)	-
AI use	Day-to-day tasks		3.9 ±1.29	
	Work	5.8 ±1.13		5.7 ±1.01
Age	18-24 years	16.7% (n=3)	80% (n=16)	20% (n=2)
	25-34 years	55.6% (n=10)	5% (n=1)	70% (n=7)
	35-44 years	16.7% (n=3)	10% (n=2)	10% (n=1)
	45-54 years	5.6% (n=1)	5% (n=1)	-
	65 years and above	5.6% (n=1)	-	-
Gender	Female	33.3% (n=6)	45% (n=9)	10% (n=9)
	Male	66.7% (n=12)	45% (n=9)	90% (n=1)
	Non-binary	-	5% (n=1)	-
	Prefer not to say	-	5% (n=1)	-

3.3.1.1. Participant Recruitment and Characteristics For the focus group workshop, the participant pool consisted of all 85 seminar attendees. Participation in the workshop was voluntary. Among the 85 participants, 16 were women, 58 were men, and 11 preferred not to disclose their gender. By the time of the focus group workshop, $N = 18$ participants remained (Table 3.1). All of them had received formal training in AI. They had prior experience in various AI sub-fields, including machine learning ($n = 75$, 90%), deep learning ($n = 63$, 76%), natural language processing ($n = 54$, 65%), video analysis ($n = 19$, 23%), and speech processing ($n = 14$, 17%).

3.3.1.2. Setting and Materials The focus group workshop was held in an auditorium to accommodate all attendees comfortably, with tables and chairs arranged in rows. The first two authors of the paper acted as facilitators of the workshop. The participants were provided with markers, pens, post-its, A4 sheets, and other materials necessary for engaging in the process.

3.3.1.3. Procedure The focus group lasted for approximately an hour and encompassed two main activities: (i) *Introduction* (ii) *Brainstorming*.

Introduction The workshop began with an introduction outlining its objectives and agenda, emphasizing the importance of their participation. Participants were familiarised with the concept of AVIs in the seminar. To acquaint them with potential feedback types for video

interviews, we used designs[27] from existing literature [14, 115], selected based on a review of automated soft skills assessment in social interactions [260]. Participants were also reminded of the techniques and algorithms explored in the seminar, pertinent for evaluating and predicting performance in video interviews. This activity was allotted 15 minutes.

Brainstorming Participants were divided into smaller groups of 3 to 4 individuals. Structured brainstorming sessions were conducted to catalyze ideation. The facilitator invited them to share their perspectives on the advantages and disadvantages of AI technologies for automated interview training and feedback.[28] Further, the discussion transitioned towards generating potential ideas for addressing the identified challenges or enhancing the benefits of AI. The participants were encouraged to reflect on their professional experiences and insights, allowing for nuanced discussions on the multifaceted nature of AI. This activity lasted around 45 minutes.

Throughout this brainstorming activity, participants were encouraged to document the potential ideas and solutions discussed using the materials provided, whether through written notes or sketches. These served as artefacts of ideas generated during the session. The two facilitators actively circulated among the participant groups, engaging in discussions and taking detailed notes. By moving fluidly between groups, the facilitators ensured that all perspectives were heard, enriching the overall breadth and depth of the data collected during the session. At the end of this activity, we collected eight solutions and recommendations (referred to as FG1, FG2,..., FG8).

3.3.2 Co-Design Workshop with Job Seekers

In phase 2, we conducted a co-design workshop with job seekers to apprise them of the potential AI-driven solutions for automated training systems and understand their needs.

3.3.2.1. Participant Recruitment and Characteristics For the co-design workshop, we utilized a participant pool from our university, comprising over 8,000 volunteers for behavioural experiments. We used the following screening criteria: (i) actively seeking employment; (ii) attended at least one job interview in the past six months; (iii) age 18 or older. They did not have prior experience using commercial hiring platforms as criteria, as our goal was to capture expectations and concerns without bias from specific products. Invitations were sent with a screener questionnaire covering demographics, interview preparation strategies, and eligibility criteria. Based on matched criteria and availability, we recruited $N = 20$ job seekers (Table 3.1). Six were part-time employed, two were full-time employed, and the rest were unemployed. Sixteen were students. Participants reported diverse preparation strategies, including advice from peers, online tips, social media (e.g., Facebook, LinkedIn), and practising common questions. The participants belonged to different faculties including business and economics ($n = 3$), social and political sciences ($n = 3$), geosciences and environment ($n = 3$), basic sciences ($n = 2$), life sciences ($n = 1$), engineering ($n = 1$), computer

[27]The designs used can be found in the supplementary. All supplementary materials of the paper are available on OSF at `https://osf.io/763eq/?view_only=46f6d3414f7b4055a7f707e3ce5dd0d4`.

[28]The questions that guided the focus group are provided in the OSF supplementary materials.

and communication sciences ($n = 1$), biology and medicine ($n = 1$), and law and criminal sciences ($n = 1$).

3.3.2.2. Setting and Materials The design workshop was held in a spacious laboratory room to accommodate all participants comfortably, with tables and chairs arranged to facilitate group collaboration. The workshop was conducted in a single session by the first two authors, who served as the facilitator and the assistant. To support the design activities, participants were provided with design kits containing markers, colour pens, post-its, A4 sheets and other materials necessary for engaging in the design process.

3.3.2.3. Procedure The workshop lasted for approximately three hours and fifteen minutes and encompassed four main activities.: (i) *Introduction* (ii) *Brainstorming* (iii) *Sketching* and (iv) *Evaluation*.

Introduction The workshop began with an explanation of its goals. Participants were then introduced to the problem of automated training and feedback for job interviews. To illustrate this concept, we demonstrated an asynchronous interview video where an interviewee responded to behavioural questions displayed on the screen. This demonstration aimed to enhance participants' understanding of video interviews and their recording process.

The solutions from the focus group were organized into an affinity diagram [32] and summarized as concepts for a proof-of-concept approach to providing video interview feedback. These concepts, including virtual interviewers and transparency in assessment, were presented to job seekers in bullet points as slides. The goal was to help them understand AI capabilities and think critically about automated interviews. To further support comprehension, feedback designs from the literature (see §3.3.1.3) were shared as examples for participants to build upon. The interactive session, lasting about 15 minutes, allowed for questions and engagement.[29]

Brainstorming Following the introduction, participants were randomly divided into groups of three members each, forming a total of seven groups (Groups A to G), with one group consisting of only two members. Each group was tasked with listing the pros and cons of the asynchronous video interview systems and feedback designs presented earlier. They discussed these points among themselves and were instructed to identify the most significant issues by marking them with a $\star$ symbol. This exercise aimed to help group members establish common ground and ideas, as they would continue to collaborate based on these ideas in subsequent stages. By the end of the activity, each group was required to collectively narrow down their list to a maximum of three issues they considered most important. This activity lasted around 25 minutes.

Sketching Following the brainstorming session, participants were introduced to sketching techniques through a brief presentation featuring examples from various contexts, such as user interfaces for learning and privacy tools. This helped them understand how sketches

[29] All materials used in the activities can be found in the OSF supplementary materials.

can be used to conceptualize technological solutions. They were then asked to conceive potential technological solutions for each issue they had previously marked with a $\star$ symbol. Participants were informed that their solutions could either enhance existing technologies (similar to the feedback examples shown in the introduction) or propose entirely new technological solutions (such as novel applications or services). Continuing to work in their groups, participants were required to sketch at least two design solutions or ideas. To facilitate the sketching process, participants were given practical tips and shown examples of sketches that were intentionally rough and unpolished. This approach was intended to reduce any anxiety about perfection and encourage creativity. Additionally, participants were asked to assign a name to each of their proposed solutions to help clarify and conceptualize their ideas more effectively. This activity lasted around 60 minutes, followed by a 20-minute break. At the end of this activity, we collected 27 design solutions with 3-5 designs from each group (A1, A2,.., B1, B2 and so on).

Mutual Feedback The primary purpose of this final activity was to collect feedback from participants on the designs. Participants presented their designs in brief flash talks lasting under three minutes each. After each presentation, participants evaluated the effectiveness of each design using an online evaluation form. They provided ratings on effectiveness and shared open-ended comments on overall impressions, the suitability of the solutions in real-life scenarios, and suggestions for further improvements. This evaluation aimed to gather diverse perspectives on the practicality and potential impact of the proposed designs.

3.3.3 Online Evaluation with AI Practitioners

In phase 3, we conducted an online evaluation session with AI practitioners to gather their expert insights on the designs proposed by job seekers. By involving AI practitioners, we aimed to align user expectations with the technical feasibility of AI solutions.

3.3.3.1. Participant Recruitment and Characteristics We employed snowball sampling for recruitment, where the participants were asked to identify other potential subjects. The initial set of practitioners ($n = 3$) was recruited through our professional networks. We aimed to match the participant characteristics with those of the focus group.[30] Hence, we used the following screening criteria: (i) has formal training in AI, statistics, or a related field, (ii) has more than 2 years of experience working with AI, and (iii) age 18 or older. Invitation emails were sent to 20 participants, and we recruited $N = 10$ who agreed to take part (Table 3.1). They had prior experience in various AI sub-fields, including machine learning ($n = 9$), deep learning ($n = 10$), natural language processing ($n = 8$), computer vision ($n = 5$), and speech processing ($n = 2$).

3.3.3.2. Procedure We developed an online survey to evaluate the collected designs, conducted in three steps. First, participants completed a demographics questionnaire. Next,

[30]The data from the focus group was collected anonymously due to privacy policies, and hence, we could not reach out to them in phase 3.

they were introduced to the problem of automated interview training and feedback, and AVIs by providing a link to an AVI. Following this, they were given a brief overview of the task, where they analyzed one design at a time and answered related questions. Close-ended questions assessed the effectiveness, feasibility, and implementation complexity of the design features, while open-ended questions invited feedback on the usefulness of the features and any suggestions or improvements. As a part of the open-ended questions, participants were encouraged to annotate specific design elements on paper and upload them to the survey.[31] Overall, we received 40 pieces of feedback on 20 designs, with each participant reviewing 2-3 designs.

3.3.4 Data Analysis

The data we collected and analyzed included the following. From phase 1: (a) 8 solutions, (b) researcher notes, From phase 2: (a) list of pros and cons of AVIs, (b) 27 designs, (c) video recordings, (d) mutual feedback; From phase 3: (a) 40 evaluation feedback on designs (b) answers to open-ended questions

A thematic analysis approach [45, 46] was employed to examine the data collected from each phase separately. Two authors independently coded the data, utilizing an inductive approach wherein codes and themes were derived from the data. The authors held regular meetings to recursively review and refine the codes throughout the coding process. Any disagreements were then jointly discussed and resolved. These codes were organized into a thematic map and grouped into sub-themes reflecting their prevalence within the data set. Ultimately, this process identified three themes in Phase 1, two overarching themes in Phase 2, and two themes in Phase 3.[32]

Finally, we examined the outcomes of all three phases together to understand the overlaps and divergences between their perceptions of automated interview training systems. Through this process, the findings can better inform future designs, bridge gaps, and harness the synergy between these stakeholder groups.

3.3.5 Ethics

All three phases of the study were approved by the authors' Institutional Review Board (IRB). Participants were provided with informed consent forms that outlined the purpose, the use of the collected data, and their rights as participants. Confidentiality and anonymity were maintained, with all data handled in accordance with established research ethics guidelines. The data collected from AI practitioners was anonymous, as we did not collect their names. Each job seeker received a monetary incentive equivalent to 65 USD for their participation in the co-design workshop. AI practitioners did not receive any monetary incentive. The research materials for this study are shared in accordance with the criteria for transparency and research ethics outlined by Salehzadeh Niksirat et al. [274].

[31] All the materials used are shared in the OSF supplementary materials.
[32] The code book used is provided in the OSF supplementary materials.

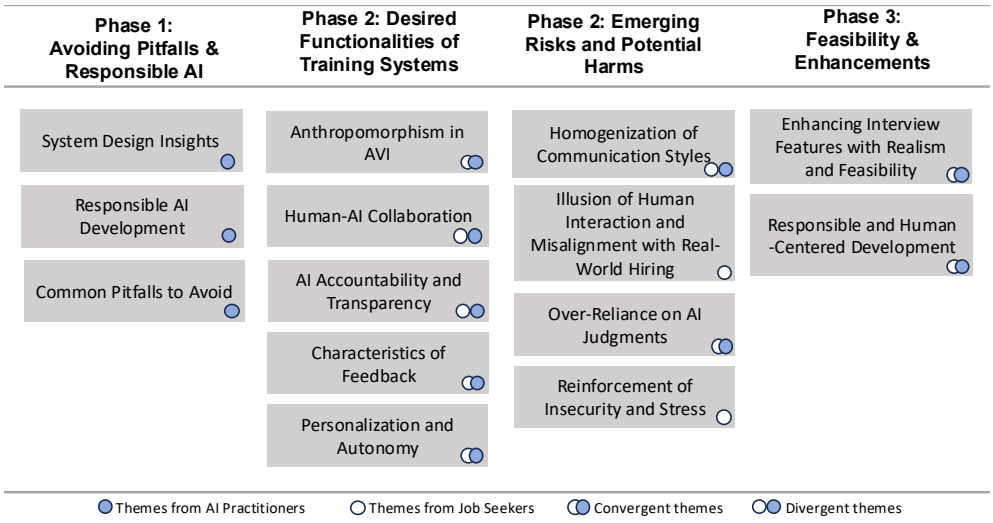


Figure 3.2: Summary of the themes we identified from both the AI practitioners and the job seekers across the three phases of the co-design study.

3.4 Findings

This section presents the results of our thematic analysis, organized by study phase. In Phase 1, three themes emerged from the solutions proposed by AI practitioners during the focus group. In Phase 2, we report two sets of themes: one based on the designs created by job seekers, and another based on the mutual feedback they provided on each other's designs, complemented by the pros and cons they identified. In Phase 3, two themes emerged from the online evaluations of these designs by AI practitioners. Across phases, we identify key patterns from each stakeholder group, indicating convergence and divergence. Themes of shared or complementary perspectives are denoted by ⊙, contrasting perspectives by ○○, practitioner-only themes by ●, and job seeker–only themes by ○. Adapting this notation from set theory [132] enables representation of alignment and gaps, informing the design of more inclusive and effective AI-assisted interview training systems.

We refer to the AI practitioner-generated solutions from Phase 1 using the prefix "FG" followed by their identifier (e.g., FG1, FG2). Job seeker designs from Phase 2 are labeled by group (A–G) and design number (e.g., A1, A2). In Phase 3, feedback from AI practitioners is identified by the design it evaluates (e.g., Eval-A1 refers to feedback on design A1). When quoting participants, we identify them by stakeholder role (job seeker: JS; AI practitioner: AI), design or solution proposed or associated group, and gender.

3.4.1 Findings from Phase 1: Avoiding Pitfalls and Responsible AI Development

Three major themes emerged from the analysis of solutions (Figure 3.3) produced by AI practitioners in the focus group.

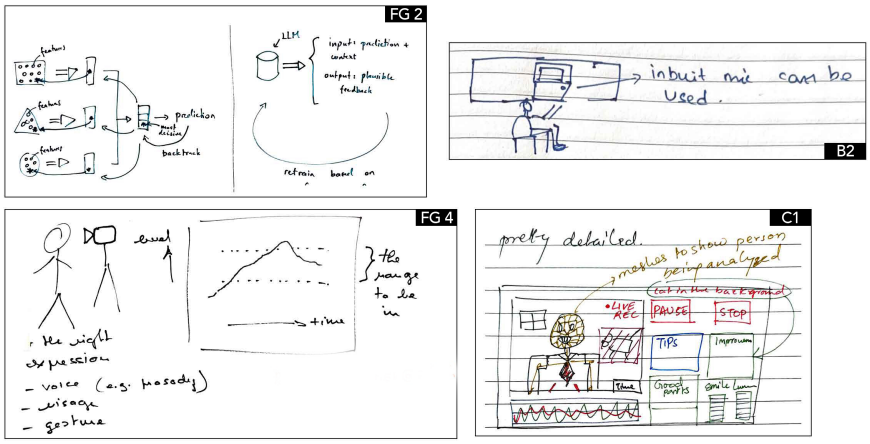


Figure 3.3: Sample sketches generated by AI practitioners in phases 1 and 3 of the study. Each sketch from phase 1 is labelled with FG and numbered. Each sketch from phase 3 is labelled by the design it was given as feedback on. The sketches illustrate various solutions to improve designs from job seekers. For example, Eval-C1 depicts the use of a face mesh to represent the feedback on facial expressions. All other feedback on designs is provided in the supplementary.

3.4.1.1. System Design Insights ○ This theme captures AI practitioners' recommendations for system features that enhance the technical robustness and training effectiveness of AVI systems. Their suggestions emphasize the importance of multimodal analysis, feedback loops, and the use of ideal performance benchmarks.

AI practitioners emphasized the importance of capturing and analyzing ***multimodal behavioural features*** in AVI systems. Many solutions proposed the inclusion of verbal and non-verbal cues such as body language, facial expressions, tone, pitch, and prosody. FG1 focused on assessing candidate impact through body language. FG4 aimed to measure voice parameters like prosody, gestures, and the appropriateness of expressions, tracking these features over time. FG5 proposed combining multiple modalities for scoring and feedback.

FG2 suggested validating AI-generated assessments through a secondary model and/or expert human feedback, creating an ***iterative feedback loop*** in which both human and model-generated feedback are integrated into the system to improve future performance. Practitioners suggested using techniques such as large language models, generative AI, and transfer learning to support this continuous learning process. *"Use transfer learning to retrain it for a specific use case."* [AI, FG2, Male]. In this setup, human feedback acts as both a corrective mechanism and a way to calibrate model outputs, enabling learning from expert judgments over time.

To facilitate skill development, many practitioners suggested incorporating ***comparison over multiple sessions or with ideal performance***. *"[The candidate] tries different sessions to reach a level of feedback."* [AI, FG4, Male]. Many solutions envisioned systems that could benchmark candidate performance against ideal performance standards (e.g., loudness, expression) and highlight deviations. *"Acceptable level versus their current level of all the multimodal features like right expressions, loudness, and pitch."* [AI, FG4, Male]. Practitioners also proposed flagging mismatches between verbal content and facial expressions: the difference in what they are saying and how they are saying it. *"Detect*

incongruence and provide feedback on this. For instance, person smiles but says something contradicting." [AI, FG3, Male]

Some solutions explored innovative ***feedback delivery formats***. FG8 proposed using AI-generated media, such as cartoon-style videos or idealized interview renderings, suggesting a multi-modal feedback approach. *"For e.g., video cartoon example of your interview style vs ideal."* [AI, FG8, Male]. FG3 echoed this: *"Provide examples of how to be/what to do to the users. Use GenAI to do it."* [AI, FG3, Male].

3.4.1.2. Common Pitfalls to Avoid ○ This theme reflects practitioners' awareness of several recurring design pitfalls and the strategies they employed to mitigate them. Their concerns included over-guiding users, overfitting user behaviour to AI metrics, and neglecting real-time adaptability in training.

One such concern was ***spoonfeeding feedback***, providing answers too directly without supporting user learning. FG1 advocated delivering feedback as coaching that prompts users toward self-discovery: *"Instead of giving feedback directly, do not tell them the answer; lead them to the answer."* [AI, FG1, Male].

Another strategy underlined was the ***balance between standardization and personalization***. FG1 emphasized using structured interviews and standardized questions to ensure fairness, but cautioned against uniformity across all domains and roles. Practitioners noted that some flexibility is necessary to account for domain-specific expectations, diverse candidate backgrounds, and the subjective nature of interviews. FG8 suggested adapting training to the specific needs and dynamics of the interviewer or teams, fine-tuning models to reflect their unique styles and feedback preferences. *"But in small teams, it is too specific to the interviewer/team. Eg, PhD Professor's communication style hiring a student."* [AI, FG8, Male].

Another pitfall involved the risk of ***overfitting user behaviour to AI evaluation metrics***. Practitioners cautioned against enforcing uniform scoring criteria across contexts. Such standardization could encourage candidates to optimize for system expectations over authentic, contextual responses. FG4 suggested multiple practice sessions to evaluate performance, so the assessment does not overfit to one sample. They also warned that candidates may learn to *game* the system with practice. FG8 proposed mitigating this by allowing customization of feedback criteria and tailoring training to the communication styles.

A ***lack of real-time adjustability*** was also highlighted as an issue. Practitioners noted that purely post-hoc feedback may limit candidates' ability to respond dynamically during interviews. FG3 and FG4 proposed live visual or textual guidance during the interview to help users adjust voice modulation or facial expressions dynamically. FG4 suggested using a combination of post-hoc and real-time feedback. *"As they practise, they get feedback and get to try another time to do better in real-time."* [AI, FG4, Male]. This design choice aims to promote more responsive, iterative learning.

3.4.1.3. Responsible AI Development ○ A prominent theme that emerged was how AI practitioners approached ethical considerations in the design and evaluation, particularly around bias mitigation, interpretability, and system transparency.

Practitioners acknowledged the challenges of ensuring fairness and inclusion across diverse user groups. While they recognized the importance of dataset diversity, their emphasis was on ***bias mitigation and inclusive modelling***. FG8 and FG5 proposed developing adaptable models tailored to specific user needs, such as introverts or people with different abilities. This included strategies like fine-tuning thresholds or creating specialized sub-models for distinct demographic or behavioural groups. However, they also acknowledged the limitations of current approaches, asking: *"How can we account for people with different neurodiverse conditions, like autism?"* [AI, FG8, Female], underscoring the complexity of building genuinely inclusive systems.

Several solutions focused on improving ***explainability and trust***. They suggested two types: to interpret predictions using the features used to train the model, or to use external sources of interpretability. Solutions FG2 and FG6 emphasized highlighting specific ***features that influenced assessment***. *"Backtracking to the most decisive feature will be useful."* [AI, FG2, Male]. Another recommendation was to involve ***temporal alignment*** between video and textual feedback. FG5 suggested synchronizing feedback with specific moments in the interview recording to see exactly where and why certain assessments were made. This could enrich learning by anchoring feedback to concrete behaviours and making system reasoning more transparent. *"While giving feedback, align the video with the text, so that feedback can be better understood."* [AI, FG5, Female]. FG6 suggested tools like SHAP values [196] to clarify feature importance, along with counterfactual explanations that show what changes could lead to a more favourable outcome [196].

AI practitioners also emphasized the need to create mechanisms for ***feedback-informed accountability***. FG2 proposed using feedback loops not only for continuous improvement but also for algorithmic auditing and monitoring. They stressed the value of documenting feedback interactions with experts to enable auditing, improve traceability, and ensure transparency in assessments. Complementing this, FG7 called for empirical user studies to evaluate the appropriateness of different features before large-scale deployment. *"How does it change when interview questions are not read by the interviewer, but it is a virtual (human) avatar conducting the interview?"* [AI, FG7, Female]. They supported adding interactive elements like avatars or simulated interviewers to enhance the interview process, but recommended validating their effectiveness and user impact in context.

3.4.2 Findings from Phase 2: Desired Functionalities of Training Systems

The analysis of the 27 design artifacts (Figure 3.4) from job seekers resulted in five themes.

3.4.2.1. Anthropomorphism in AVI ◎ Anthropomorphism is the attribution of human qualities, such as behaviours, intentions, and emotions, to non-human entities [246]. Job seekers often projected human interviewer capabilities onto AVI systems. They desired key human-like elements, such as conversational interviews and inclusion of uniquely human emotions, such as empathy. Similar to practitioners in Phase 1, job seekers echoed support for interactive features, while also extending these expectations to more affective qualities.

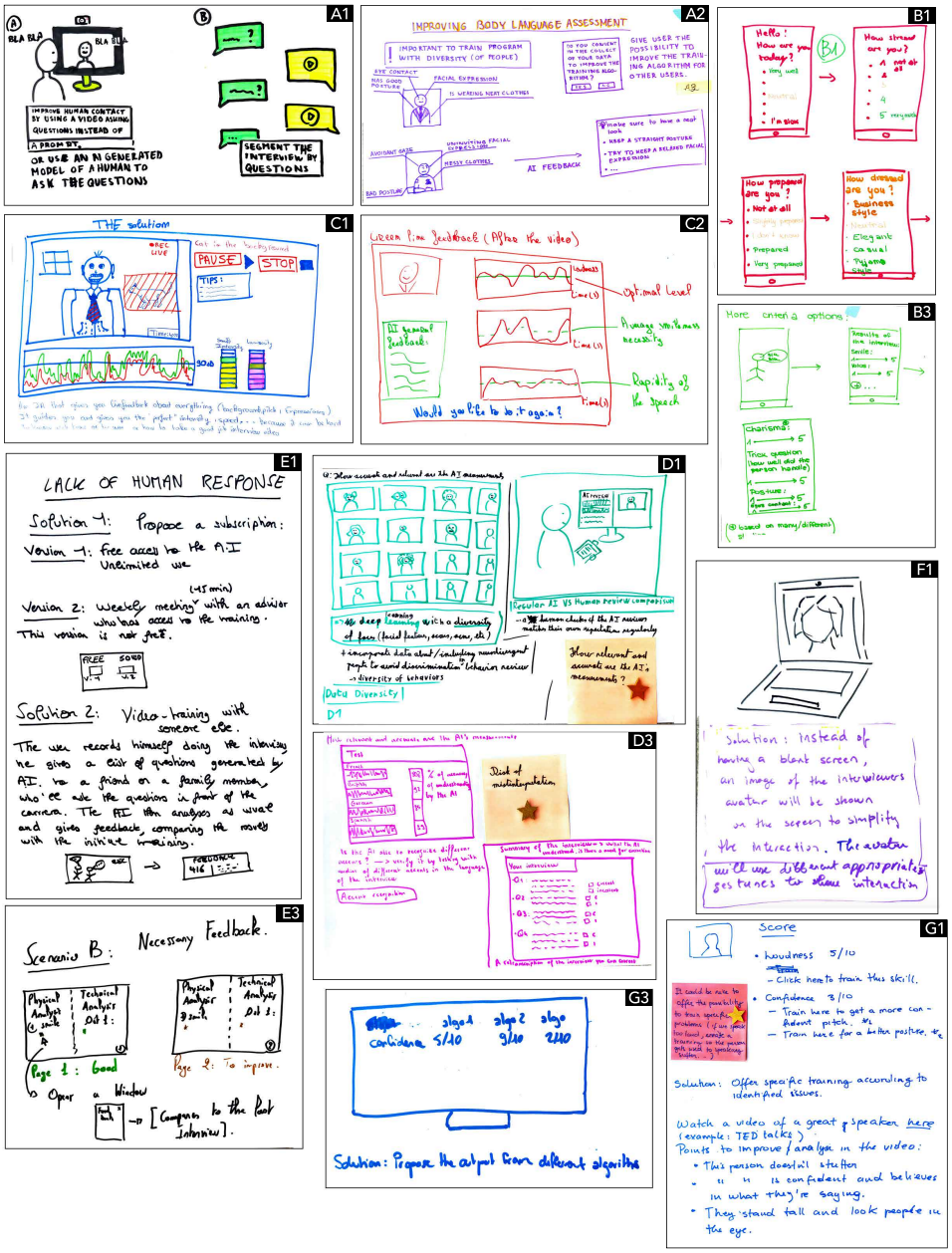


Figure 3.4: Sample sketches generated by job seekers in phase 2 of the co-design study. Each design is labelled with groups A through G and numbered within each group(e.g., A1, A2...), The designs illustrate various solutions to improve interview training systems like conversational video chat-like system with virtual interviewers (A1, F1), feedback on optimal levels of parameters like facial expressions, loudness, charisma (A2, C1, C2, B3), human validation of AI-assessment (D1, D3, E1), risk of misinterpretation (B1, D3, G3) and personalized training (B3, C2, G1). All 27 designs are provided in the supplementary in OSF .

Many designs had job seekers ***envisioning a conversation exchange*** with a virtual counterpart rather than a monologue answering pre-established interview questions. For instance, A1 depicted a real-time, chat-like interface for exchanging video conversations. *"Training with AI isn't like training with humans, so a solution is to have the video of*

someone or a recruiter asking the question and give time to answer, submitting one video at a time." [JS, A1, Female]. Job seekers expressed a preference for follow-up questions contextualized based on their responses to maintain the flow of conversation in A3 and F3. They preferred receiving explanations or hints when encountering difficulties while answering, mimicking real-world scenarios. In addition, some job seekers preferred to receive feedback through a chat format rather than a static display.

Designs A1 and F1 incorporated either ***virtual humans or videos of humans asking questions***. Most job seekers preferred this approach over reading questions from a screen prompt, as it more closely mimics a real interview. "*In a virtual interview, we will speak to the camera, and there is no person. So it would be good to have the avatar of the interviewer mimicking the emotions and expressions of a real person.*" [JS, F1, Female]. Designs further enhanced the experience by using virtual reality, creating an immersive environment where candidates could interact with a virtual interviewer.

Designs B1 and E2 advocated for ***incorporating candidates' emotional, physical, and psychological states*** into the interview process. They endorsed a system characterized by empathy that considers a candidate's circumstances. They proposed employing a pre-interview questionnaire to evaluate the state of mind and well-being of candidates, and subsequently adapting the interview: *"AI never considers what you feel, asking something like 'how are you feeling?', 'what happened yesterday and what do you feel about it?' can help with the feeling in the job interview."* [JS, E2, Male]. This approach aims to cultivate a supportive and understanding interview environment, recognizing the potential impact of candidates' well-being on their performance and experience.

3.4.2.2. Human-AI Collaboration ∞ A prominent theme that emerged from the job seekers was human-AI collaboration. Job seekers emphasized the need for humans and AI to work together, enhancing rather than AI replacing human capabilities. Their suggestions encompassed human validation of AI assessment and AI-augmented training.

Many designs advocated for ***human validation of AI assessments***. Job seekers suggested that regular or random checks of AI-generated assessments by experts are essential to maintain trust and expected standards. Two specific methods were proposed: first, an expert could randomly assess a video interview independent of the AI's assessment to measure accuracy; second, an expert could ***regularly review AI assessments*** to ensure alignment with expectations. *"Have a human review random interviews so as to match the expectations."* [JS, D1, Female]. They also suggested allowing candidates to ***review and correct the transcripts*** of their recorded videos. *"Approve or disapprove and correct what you said and what you did not."* [JS, D3, Female]. They viewed the transcripts as a proxy for understanding the system's interpretation of their responses.

Two solutions from design E1 (Figure 3.4, E1a and E1b) emphasized the ***role of experts in AI training***. E1a suggested a subscription model where candidates have unlimited access to AI-based training, with an option for paid human expert training once a certain proficiency level is reached. *"Start by yourself by recording and getting [feedback] from AI... after you think you are ready to have an actual [video] meeting with a person with [access to] videos you already did to point at improvements. Learn from AI experience and at the same time human experience."* [JS, E1, Male]. In this scenario, the human expert would

have access to comprehensive AI feedback over the training period, allowing for in-depth and personalized feedback. E1b proposed an alternative approach where a ***peer would ask AI-generated questions*** and record responses for feedback through the system. Although the system would still conduct the analysis, this method introduces a human collaborating with the system, enhancing the training experience.

Unlike Phase 1, where AI practitioners primarily framed human involvement as a way to improve system performance, job seekers emphasized collaboration that validated assessment and safeguarded their agency. They stressed reviewing transcripts, validating assessments, and integrating human feedback directly into their training process.

3.4.2.3. AI Accountability and Transparency ∞ Accountability and transparency emerged as central themes in discussions among job seekers. Major sub-themes were transparency and risk of misinterpretation.

Job seekers, like practitioners, also valued transparency in the feedback process. Several designs suggested using a ***timeline that links feedback to specific parts of the interview video***. *"After the interview, red points can be placed on the timeline. Specific explanations of red points... correspond to the most significant errors."* [JS, C4, Male]. They liked having the ability to self-review the video alongside the feedback timeline, ensuring clear reasoning behind AI decisions. *"Algorithm should show the user when or where they did something wrong."* [JS, A4, Female]

Job seekers emphasized the need for accountability beyond transparency, advocating for mechanisms to validate AI decision-making and ensure fairness, while practitioners had mainly focused on transparency through explanations and interpretability. Their concerns about misinterpretation, bias, and over-reliance led to design proposals such as accent recognition, transcript correction, multiple algorithms, and multimodal baselines to make AI assessments contestable and verifiable in practice.

Several designs, including B1, D3, G3, F2 raised concerns about the ***risk of misinterpretation***. For example, D3 proposed an *accent recognition* feature to ensure the system could accurately interpret diverse accents: *"Test AI with different audios of the same text but with different accents."* [JS, D3, Female]. D3 also suggested allowing candidates to correct any misinterpretations in the speech transcript.

Job seekers called for using multiple algorithms for assessment to ***prevent over-reliance on a single model*** and mitigate bias and errors. *"The output may depend a lot on the algorithm and the training set. It would be good to add the output of different algorithms in the feedback.... to give the idea of average feedback."* [JS, G3, Female]. Additionally, B1 proposed capturing qualities like the candidate's voice, facial expressions through a ***repeat after me feature***, where candidates would recite standard sentences to establish a multimodal baseline for assessment. F2 suggested using smart cables and external hardware to monitor physiological responses during the interview, as a way to enhance the precision and accuracy of the AI's assessment.

3.4.2.4. Characteristics of Feedback ○

Characteristics of feedback is the most discussed theme among job seekers. It mainly addresses the specific qualities and features that make feedback effective for performance

improvement as per the job seekers' needs. It comprised multimodal features for feedback, training for each job parameter with granular and actionable feedback, which echoed practitioners' focus on capturing multimodal behavioural features while also extending expectations toward greater granularity and usefulness.

Several designs, like A2, B3 and G1, analysed low-level features like posture, eye contact, smile, and loudness, as well as high-level qualities like charisma, confidence, and communication. *"AI can take note of body language like appearance, eye contact, if you are looking directly into the camera, if you have a good posture, facial expressions, etc.,"* [JS, A2, Male].

Job seekers emphasized the need for ***granularity of feedback***. Most designs proposed dividing feedback into sections, for example, by job-related parameters like charisma or by interview questions, or by separating non-verbal from content-related feedback. *"More the feedback [in different parameters], more precise the details will be."* [JS, B4, Female]. *"In each category, have a sub-category [of feedback] for different aspects."* [JS, D4, Female]. B3 pointed out that parameters like charisma might be challenging to assess and should be supported by literature. They suggested using questions specifically designed for measuring such parameters. *"Charisma.. might be difficult, but based on many different studies, it might be possible to check if the person is charismatic or not depending on the position."* [JS, B3, Female].

Several designs suggested offering targeted strategies to improve specific skills. They emphasized the importance of ***training for each job-related parameter***, optionally including the candidate's progress over multiple sessions, with separate and detailed feedback for each area. For example, training for communication skills, confidence, charisma, etc., separately, depending on performance on each. *"Under each category, for example, confidence - click here to get more specific feedback, and it would take us to another page [for training], which would say, for example, here is a video of a great speaker. You can analyse the video. This person talks with a lot of confidence, he speaks loud enough, makes eye contact, etc.,"* [JS, G1, Female]. B4 and C2 suggested that training should offer clear ***direction of fine-tuning each attribute***, helping individuals work toward optimal performance in a structured, step-by-step manner. *"Lot of people may not understand what the feedback was supposed to mean.... So here we have the graph for loudness [where] the green line is the optimal decibel level that needs to be followed."* [JS, C2, Male].

Several designs underscored the need for ***actionable and timely feedback***. A common concern among job seekers was how to make feedback truly useful for improving performance. As one participant noted, *"Feedback gives us a score, what can we do about it?"* [JS, G1, Female]. The majority of designs proposed various ways to make feedback more actionable, such as using ***ideal video examples to illustrate optimal performance*** and offering comparative analysis with past interviews. ***Real-time feedback*** was a key feature, with C1 and D2 suggesting live graphical comparisons of optimal versus actual performance across features or live textual guidance to help candidates adjust aspects like voice modulation or facial expressions. For instance, ***visual or textual cues could indicate where to adjust*** tone or facial expressions during the interview. As one participant mentioned, *"In this live feedback, you have access to all parameters' feedback so you can improve yourself in the interview and doing more videos."* [JS, C1, Male]. These features would enable job seekers to apply feedback effectively and continuously enhance their performance.

Design C3 resonated with the suggestion of AI practitioners using GenAI, where they suggested rephrasing the candidate's answers optimally as an ideal version of the recorded answer. *"Give you [the answer on] how you can change what you said, for it to be a little more formal or professional."* [JS, C3, Female]. Design B4, too, supported this, saying that each session should include an objective or a mission to be achieved.

3.4.2.5. Personalization and Autonomy

Personalization and autonomy are key in designing AI-assisted training systems, especially for job interview preparation, where user needs vary widely. Previous literature indicates the characteristics of choice and personalization are associated with autonomy [257, 266]. This was also a point of discussion among the job seekers in the study. This theme showed alignment on the importance of personalization, but differed slightly in emphasis: practitioners focused on system-level adaptation for interviewers and teams, job seekers elaborated on it, and stressed user-level autonomy and control in training.

Section 3.4.2.4 discussed personalization through tailored feedback and custom learning paths through targeted training. Job seekers expanded on personalization by ***tailoring interview content to specific needs*** in several designs. Design C3 suggested determining the field or industry the candidate is applying for and tailoring training and feedback. *"Check which fields we are trying to get a job, so AI can know from which database to get the vocabulary."* [JS, C3, Female]. C3 and F3 also proposed analyzing how much industry-specific vocabulary is used to assess a candidate's expertise. Additionally, A2 recommended allowing candidates to ***donate their own videos to further personalize the AI assessment***, ensuring the system adapts to their individual communication style and skills, giving a more accurate assessment.

Job seekers further emphasized ***autonomy in training*** settings. Design B2 proposed customizable interview settings, allowing candidates to choose between virtual or human interviewers and specify their gender, number, style, mood, and difficulty level, including the option to simulate stress interviews if necessary for the upcoming interview. While many job seekers wanted feedback compared to past performances, E3 suggested showing comparisons only when requested.

3.4.3 Findings from Phase 2: Emerging Risks and Potential Harms

The analysis of mutual feedback provided by job seekers on each other's designs and the list of pros and cons revealed that job seekers, despite seeing the advantages, identified certain emerging risks and harms of AI-assisted interview training systems. This resulted in four potential emerging risks as perceived by job seekers.

3.4.3.1. Homogenization of Communication Styles

One theme that emerged strongly from the job seekers' feedback and design discussions was the homogenization of communication styles in AI-driven interview training. Many participants expressed concern that such systems risk promoting one-size-fits-all communication norms, leading to ***unnatural, rehearsed behaviours over time***. *"A norm [assessment criteria] doesn't really take into account the differences between people and the technical defects of the material used."* [JS,

A, Non-binary]. Repeated exposure to fixed feedback loops was seen as ***overfitting to AI feedback*** and encouraging over-coaching. Job seekers feared shaping their responses to align with what AI rewards can reduce authenticity and spontaneity. *"Difficult in real life as the repeating aspect of answers generated by AI by the seeker could be evasive for interviewer."* [JS, E, Male]. One participant described this outcome succinctly as: *"Robotic answers"* [JS, E, Male].

Participants noted that ***context-blind evaluations***, that do not adapt to role-specific, cultural, or interpersonal nuances, can misrepresent candidate potential and disproportionately disadvantage those with diverse communication styles. For instance, traits like charisma, which are deeply personal and culturally contingent, were viewed as being oversimplified when reduced to quantifiable metrics. *"It sounds a little too subjective for evaluating charisma. It would need to develop a very objective evaluation of charisma."* [JS, D, Female].

Several participants highlighted accent bias and linguistic insensitivity, noting that systems that fail to account for non-native fluency, code-switching, or regional dialects may unfairly penalize candidates. *"It should take into account if the language that we are speaking is our mother tongue and therefore our vocabulary cannot be that profound."* [JS, B, Female]. The ***erosion of cultural and linguistic diversity*** was seen as a significant risk *"Can it take other languages' words into account, as some languages mix sometimes in a speech?"* [JS, A, NB]. Across these discussions, many participants viewed human oversight as essential to preserving variation, contextual understanding, and authenticity in the interview process.

Both practitioners and job seekers cautioned against overfitting users to AI-driven norms, but framed the concern differently. Practitioners highlighted design pitfalls such as spoon-feeding feedback, overfitting behaviour to AI metrics, while job seekers framed it as a socio-psychological risk, warning that homogenization and robotic behaviours could erode individuality and authenticity over time.

3.4.3.2. Illusion of Human Interaction and Misalignment with Real-World Hiring ○ A prominent concern among job seekers was the gap between simulated interview experiences and the realities of real-world hiring. Participants often critiqued the ***illusion of social presence*** devised by AI-assisted training systems. *"A lot of people might do that [AI-generated videos] to convince themselves but we truly know that those people are not real so it will be really different than in real life."* [JS, C, Male]. This created a risk of disillusionment as users realized that AI couldn't match human interviewers' empathy, emotional pacing, or the ability to adjust based on tone. *"In real life, we have empathy of the recruiter."* [JS, B, Female].

Several participants discussed the ***over-anthropomorphization of AI***, where systems create unrealistic expectations of human-like behavior, only to fall short in practice. *"I wouldn't use AI-generated videos for this because I think they might feel too uncanny and work against the intended goal."* [JS, D, N].[33] When there is a mismatch between the visual realism of AI-generated avatars and their communicative nuance, it can result in discomfort

[33]The participant preferred not to disclose the gender.

or misplaced trust in the system's capabilities. *"This will not change the fact that we speak to a screen."* [JS, C, Male].

Participants also raised concerns about ***false equivalence between AI-based training and actual interviews***. They noted the risk of optimizing responses for artificial cues, such as facial tracking or voice modulation scores, that do not translate into real hiring contexts. *"Simulating the job interview in front of the camera could not really replace the real-life interview."* [JS, E, Male]. Some highlighted how AI systems ***overlook emotional nuance and spontaneity***, flattening interpersonal dynamics and relying on limited biometric signals to infer complex traits. *"I'm not sure how to make sure the AI analyzes someone's 'humanness' correctly."* [JS, D, N]. In this context, participants repeatedly emphasized human oversight to ground AI systems in hiring realities. *"Good idea, implement AI to work with human. I strongly agree, you will be more prepared with this."* [JS, A, Male].

Finally, participants expressed doubt about the ***real-world transferability*** of training outcomes. While AI systems might improve structured responses, they were seen as inadequate for preparing candidates to handle stress, uncertainty, and improvisation in real interviews. *"I didn't really understand the point of stress when you will never achieve the same level at home with a computer than in the real situation."* [JS, A, Male]. Systems that set unrealistic expectations were viewed as potentially misleading and detached from hiring ecosystems. *"Companies won't ask you to act in a certain way, the purpose of the interview is to see how you act in this situation."* [JS, C, Male].

3.4.3.3. Over-Reliance on AI Judgments Another emerging theme from job seekers' reflections centered on the risks of over-relying on AI systems for training and assessment, especially when such systems offer a false sense of control, accuracy, or personalization. Participants also raised issues around transparency and trust calibration. While features like editable transcripts were valued for self-correction, some worried that such transparency could foster ***overconfidence in AI feedback***, leading users to misjudge their readiness or overlook human-centric aspects of performance. "Too much confidence about the answers generated by AI in practical training for job interview." [JS, E, Male].

Job seekers expressed concern that ***oversimplification of evaluation*** could lead to users *gaming* the system rather than training authentically. They were also concerned about using superficial features like pre-interview surveys for evaluation, which could be easily manipulated to present themselves in a more favorable light. *"The idea of the survey at the beginning could benefit some seekers by lying."* [JS, F, Male].

Some participants questioned how ***re-recording interviews might set unrealistic expectations***, as real interviews offer no second chances. *"In real life, no, because you cannot repeat the interview."* [JS, B, Female]. They cautioned that reliance on iterative perfection could misalign preparation and urged: *"Good to not get lost in AI trap"* [JS, B, Male]. In this context, peer practice and human interaction were seen as more authentic and effective. *"Maybe put two people in contact doing interviews for one another, that way it is a real person."* [JS, E, Male]. Several participants advocated for social co-learning through mock interviews with real people to better simulate interview conditions and provide comparative feedback. *"Practice with more people on the same topic to compare the answers generated by AI and improve them."* [JS, E, Male].

Practitioners had already hinted at this concern in Phase 1, warning that candidates might overfit their behaviour to AI metrics or learn to game the system with repeated practice. Job seekers in Phase 2 extended these concerns further, framing over-reliance on AI judgments as a broader issue of misplaced trust, unrealistic expectations, and oversimplified evaluations that could undermine authentic preparation.

3.4.3.4. Reinforcement of Insecurity and Stress ○ Job seekers expressed nuanced perspectives on how AI-based interview training systems might reinforce feelings of insecurity, anxiety, or performance pressure, particularly through the design of real-time feedback and control mechanisms. Such features were intended to support improvement but were often perceived as ***sources of stress and insecurity***.

A common concern was the ***impact of real-time feedback during performance***. Many cautioned that immediate feedback could be distracting, overwhelming, or anxiety-inducing, especially during high-stakes simulations. *"During a real interview, the interviewers don't tell you constantly what to change, so maybe that will only add unnecessary stress."* [JS, F, Male] one noted. Others echoed this concern, noting that it could increase stress. *"I think it might be a bit confusing to have the feedback when doing the exercise."* [JS, G, Female]. *"I don't think it's a good idea because it can cause more stress to the candidate."* [JS, F, Male] At the same time, some participants acknowledged that without such feedback, users may repeat the same mistakes, leading to inefficient practice cycles and heightened pressure to perfect their performance.

While some participants appreciated the flexibility to review and rerecord interviews, many noted that this could provoke ***perfectionist tendencies*** and result in excessive rehearsal. Participants also cautioned about physiological monitoring, such as stress or facial expression tracking. They noted that it could further intensify self-consciousness during an already stressful task. *"I think it would stress some people but it could help others."* [JS, D, Female]. Finally, they emphasized the importance of balanced feedback, arguing that focusing only on errors, without positives, could be demotivating: *"Pointing out only the bad can be exhausting for the user."* [JS, E, Male].

Together, these concerns highlight a key tension: systems designed to support improvement must also avoid inadvertently reinforcing the very insecurities and stress they aim to reduce.

3.4.4 Findings from Phase 3: Feasibility and Enhancements

The analysis of feedback from AI practitioners on job seeker designs revealed two major themes.

3.4.4.1. Enhancing Interview Features with Realism and Feasibility ◎ Practitioners saw many of the job seekers' designs as emphasizing that automated interviews feel both closer to real-world interviews and providing feedback that is genuinely useful for preparation. They understood the designs as efforts to make AVIs more realistic while also checking how feasible these features would be to implement.

One main sub-theme was ***enhancing the interactivity*** needs of job seekers with diverse, richer question dynamics, where prompts build on prior answers rather than appearing in isolation. "Question diversity (use a mix of questions, technical HR, etc.), subsequent questions should be a follow-up if the question is from the previous one." [AI, Eval-A1, Female]. They further enhanced the designs, making the virtual interviewer's expressions responsive to job seekers. *"Avatar speech and facial expression changes based on the job seeker answer".* [AI, Eval-F1, Male].

Another recurring focus was ***multimodal and real-time feedback***, extending job seekers' requests for feedback on delivery and timing. Practitioners suggested metrics on answer relevance, conciseness, and body language to help candidates manage performance under time pressure and adapt delivery: *"Also keen to know about the time taken for each answer, so that the interviewee can work around that."* [AI, Eval-B3, Male]. Some practitioners suggested innovative improvements. For example, video generation methods can be used to generate an ideal version of the candidate's performance to improve training and engagement. *"Use AI-vision methods to show them a corrected version of the current video/photo, e.g., fixed [corrected] gaze, expression, etc."* [AI, Eval-A2, Male].

Personalization was also emphasized, with proposals linking analysis to candidate data from LinkedIn or GitHub: *"I would like to add an option for users to upload their CV and add their LinkedIn/Github profiles to further optimize the vocabulary."* [AI, Eval-C3, Male]. Others proposed adaptive scoring mechanisms that account for external factors like noise: *"A dynamic threshold (instead of static 90db) to score the intensity. The threshold can depend on the background noise"* [AI, Eval-C1, Male]. Inspired by design B2, they proposed industry-specific customization and options to select interviewers from different nationalities (e.g., Chinese, British, American) to suit cultural preferences.

While supporting these ideas, practitioners also reflected on ***technical feasibility*** of job seeker design proposals. Some proposals, such as empathy or charisma detection, were seen as overly ambitious: *"Multimodal analysis [for empathy] is easy to say, tough to achieve. It is a useful feature but complex to implement."* [AI, Eval-E2, Male]. Practitioners further recommended using multimodal techniques to analyse the feelings of candidates along with questionnaires, as relying solely on self-reporting may not provide a complete picture. They also critiqued hardware-heavy solutions, recommending simpler integrated approaches using computer vision: *"Instead of measuring through the devices, use computer vision to recognize the emotions and act accordingly."* [AI, Eval-F2, Male].

3.4.4.2. Responsible and Human-Centered Development Job seekers in Phase 2 emphasized trust, empathy, and fairness in their designs, and practitioners carried this forward by reflecting on transparency, accessibility, and user agency. A central concern was ***transparency and consent***, with calls for clearer communication on data handling: *"Add option for user to read about how the data will be used, who stores the data, etc."* [AI, Eval-A2, Male].

Several practitioners also underscored the value of ***accessible and empathetic interaction***, suggesting conversational interfaces that respond to user input: *"The interface can be more interactive using chat chat-like interface. The person to be given some reply to show empathy, depending on the choice selected"* [AI, Eval-B1, Male]. At the same time,

they noted the risks of relying solely on self-assessment, which may not capture job seekers' actual state: *"Pre-interview questionnaire can be slightly misleading. Users might not know how to score their feeling"* [AI, Eval-B1, Male].

Another important thread was ***leveraging user agency to improve AI model performance***. Practitioners framed features such as editable transcripts not only as candidate support but as opportunities for the system to learn and adapt: *"Show users the transcript, and let them correct individual words that AI misunderstood. Then retrain those words given the user's speech"* [AI, Eval-D3, Male]. They viewed features such as allowing job seekers to compare or average multiple algorithmic evaluations (Eval-G3) and request clarifications during analysis (Eval-A3) as valuable inputs for model fine-tuning.

Together, these reflections show that while job seekers emphasized agency to better control their preparation, practitioners reframed it within a human-in-the-loop approach. In their view, transparency, empathy, and opportunities for candidate correction were not only valuable for supporting job seekers but also crucial for continuously improving the accuracy and adaptability of the AVI models.

3.5 Discussion

The participants in our multi-phase co-design study designed and progressively enhanced 27 designs and eight solutions. Our findings offer insights into the concerns and risks job seekers perceive in automated interview training systems, their desired needs and functionalities in such systems, and expert insights from AI practitioners on the capabilities and responsible development of such systems. Furthermore, we identified convergences and divergences between the two stakeholder groups. Specifically, job seekers and AI practitioners converged on desired functionalities like anthropomorphism, actionable feedback, personalization, and ethical considerations. However, we also found divergences in two critical aspects—human-AI collaboration and accountability.

In the following, we discuss these findings and provide design considerations for the future design of automated interview training systems and training systems in general. We conclude by discussing the limitations of the study.

3.5.1 Alignment on Human-Centered AI for Interview Training

The convergence between job seekers and AI practitioners on the themes of anthropomorphism, ethical considerations of AI assessment, characteristics of feedback, and personalization and autonomy highlights the shared vision of developing more human-centered and adaptive automated interview training systems.

Both groups recognize the potential of anthropomorphized interfaces—through interactive, conversational interviews or the use of virtual humans—to create more engaging and realistic experiences. Job seekers emphasised the importance of natural, context-based dialogue for enhancing realism. AI practitioners, while aligning with this goal, recommended leveraging AI-generated media to enrich feedback and make interactions more immersive. This shared emphasis on human-like interaction could reduce the friction between human

and machine communication, making the training process more intuitive and less mechanical [163]. Our results contrast previous studies on AI agents for other purposes, like online learning [246, 322], where participants did not prefer human-like characteristics. However, anthropomorphising can mistakenly lead to viewing these systems as inherently trustworthy. This misperception can misrepresent the actual trustworthiness of the system, leading to over-reliance [246]. Hence, it is crucial to achieve a balance in integrating human-like features into AI systems by carefully designing them to enhance engagement and user experience without fostering an unwarranted sense of trust or dependency. Our contrasting findings raise the question: can these anthropomorphised features serve as a bridge for more interactive and realistic human-AI communication in training systems? And if so, how can we ensure they do not foster an inaccurate perception of the AI's understanding or reliability, leading to misplaced trust? Future research should involve job seekers to analyze these variables in high-risk contexts such as interview training and recruitment.

On the ethical front, both job seekers and AI practitioners recognised the importance of ensuring data diversity in AI models for interview training and mitigating algorithmic bias, though their emphasis differed slightly. Job seekers were particularly interested in creating inclusive datasets that account for factors like cultural background, age, and abilities. Their advocacy for features like donating videos to expand and diversify the dataset points to a solution-driven approach with a strong desire for fairness. While AI practitioners recognized the importance of data diversity, they prioritized bias mitigation through solutions like benchmarking models to detect and reduce stereotyping. Furthermore, participants acknowledged the challenges in acquiring diverse datasets and highlighted the complexity involved in comprehensively addressing bias [7, 36, 79, 89, 299]. This reveals the nuances of stakeholder engagement in AI design: while job seekers demand more data representation, AI practitioners aim for technical robustness. Both groups, however, recognise that fairness and inclusivity must be at the core of any AI-driven training system. This emerging consensus is pivotal for future developments in the field, pushing for technologies that should not only support diverse populations but also actively mitigate inherent biases.

The focus on feedback and personalisation further reveals a strong alignment in expectations from both job seekers and AI practitioners. Job seekers sought detailed, actionable, and granular feedback in line with previous work [94, 257]. Particular emphasis was on multimodal inputs—such as posture, eye contact, and verbal content to improve specific job-related skills, which have been the focus of several studies in the literature [135, 219, 224, 256, 261]. Both groups recognized the importance of tailored training sessions and allowing autonomy in the interview process. They suggested options for customizing interview settings to suit job seekers' needs and incorporating candidates' professional profiles into the training process. This demonstrates the shared understanding that effective AI systems must be both adaptive and user-centered, providing targeted, actionable insights.

3.5.2 Navigating Divergent Perspectives between Job Seekers and AI Practitioners

The themes of accountability, transparency, and human–AI collaboration revealed important divergences in priorities between job seekers and AI practitioners.

Job seekers emphasised the need for strong human oversight to prevent risks of misinterpretation at every stage of the system pipeline. They called for control over both inputs and outputs, for example, by reviewing automatic speech recognition (ASR) transcripts, validating interpretations, and adding human review loops to AI-generated feedback. These proposals reflect a precautionary stance, where accountability mechanisms are embedded from the very beginning rather than retrofitted afterwards. This resonates with prior findings that job seekers often prefer a human-in-control approach in automated hiring and training systems [175], and further illustrates a trust gap [154] that motivates them to foreground fairness and reliability over efficiency.

By contrast, AI practitioners initially framed human involvement in terms of human-in-the-loop training, where expert feedback is used to refine and improve AI models. Their focus was on scalability and technical transparency, seeking to minimise continuous human supervision in favour of explanations and interpretability features that make the system's functioning clearer. From this perspective, reducing human intervention was seen as essential for broad accessibility and cost efficiency.

Both groups conceptualized human involvement in fundamentally different ways. For job seekers, being *human-in-control* meant retaining direct authority over how the system interprets their responses, such as correcting misinterpretations, reviewing and modifying AI-generated feedback, and having clear mechanisms to challenge or override system outputs. In contrast, AI practitioners described a *human-in-the-loop* model, where human experts intervene primarily during model development, such as providing expert training data, validating system performance, or refining prompts, but not during individual assessments. In this framing, practitioners maintain control over system-level behavior, whereas job seekers seek control over instance-level decisions. These differences also shape contestability [10]: job seekers expect actionable pathways to dispute or amend evaluations, while practitioners emphasize transparency and interpretability features rather than explicit avenues for challenging outputs.

However, as Wieringa [333] highlights, transparency can be considered passive—offering insights into how the system works (i.e., '*see for yourself how it works*') without explaining why it was deemed appropriate or desirable in the first place (i.e., '*let me tell you how it works and why*'). Job seekers' emphasis on active accountability reflects their desire for mechanisms that not only reveal the inner workings of the system but also allow for validating and contesting AI assessment. Accountability is defined as an obligation to inform about and justify one's conduct to an authority [43, 228]. For accountability to be achieved, it is not enough to simply make the AI system's processes visible; there must be clear pathways for reviewing, auditing, and adjusting the system's outputs to ensure fairness, mitigate biases, and correct misinterpretations, particularly in high-risk environments like job interviews.

Notably, however, the evaluation phase of our study showed practitioners moving closer to job seekers' position: they acknowledged the value of human checkpoints at earlier stages, such as reviewing ASR outputs or validating sensitive assessments. This progressive alignment across phases demonstrates the advantage of our sequential study design, where presenting stakeholder perspectives iteratively helped surface divergences and then create opportunities for reconciliation and shared design priorities.

To address these disparities, future training systems could integrate both transparency

and accountability, incorporating user-driven corrections as a form of human validation. For example, job seekers could be allowed to correct AI-generated transcripts. Such approaches would also facilitate a *collaborative feedback loop*, where AI systems learn from users' input, refining their accuracy and interpretative capabilities. This idea aligns with ongoing research on human-AI collaboration, where humans help intelligent agents to improve the agents' performance [165, 179, 220]. When job seekers understand that their corrections are actively helping the AI's learning process, this collaborative exchange can reduce the trust gap [154, 164], promote both user well-being [204, 331, 353] and system reliability. This vision of human-AI collaboration extends beyond traditional AI support, positioning the AI as both a helper to users and a system open to help from users—a reciprocal relationship that enriches human well-being and user engagement while advancing the AI's capabilities [204, 237, 331]. This mutually beneficial interaction could further enhance job seekers' motivation to engage in training, as they see tangible improvements in their skills and feel valued in contributing to the system's development. Increased engagement in this form of human-AI collaboration holds promise for helping job seekers maintain long-term motivation to interview preparation, thereby supporting job seekers in overcoming employment challenges [257].

Furthermore, job seekers could be allowed to not only correct AI-generated transcripts but also refine responses to the same question in iterative sessions, moving towards improved performance. This history or replay of answers would enable them to self-reflect on the evolution of their answers and the improvement of their performance over time [340]. From an AI perspective, this can support transparency and potentially offer counterfactual explanations [321] on improving assessments, ultimately integrating transparency and accountability.

3.5.3 Collaborative Synergies between Stakeholders for Effective AI Training Systems

Our multi-phase co-design approach underscores the mutual benefits of actively engaging diverse stakeholder groups of job seekers and AI practitioners in the design process of automated interview training systems. This collaboration enabled a balanced approach that integrates real-world user needs with the technical feasibility insights of AI practitioners. Through iterative exchanges, both groups contributed to design refinements that not only addressed system functionality but also surfaced critical human-centered concerns that might otherwise have been overlooked.

In particular, the co-design sessions revealed job seekers' perceptions of socio-psychological risks embedded in AI interview training, such as the homogenization of communication, reinforcement of stress, and overreliance on AI. These risks extend beyond technical failures to the lived experiences of users, underscoring why participatory methods are essential for identifying harms that are difficult to anticipate from a purely technical standpoint. By voicing these concerns, job seekers ensured that system designs accounted for the broader psychological and social dimensions of fairness, trust, and usability. AI practitioners, meanwhile, brought interpretability strategies and responsible design considerations that helped situate these user concerns within feasible system architectures. This reciprocal engagement not only improved design outcomes but also supported a more responsible approach to AI development [200, 324].

Additionally, ensuring that automated training systems align with regulatory frameworks, such as the EU AI Act [73] and the U.S. AI Bill of Rights [145], is critical, especially as these applications are often categorized as high-risk AI systems by these regulatory frameworks. Engaging multiple stakeholders from the design phase supports regulatory compliance by embedding fairness, accountability, and transparency into the system. By collaborating on these critical design elements, stakeholders can preemptively align the system with regulatory standards, fostering both adherence to legal frameworks and public trust in AI-driven assessment tools.

3.5.4 Design Considerations for Automated Training Systems

We present a set of design considerations for automated interview training systems based on our findings. These principles have broader relevance to automated training systems in general for various other skills such as public speaking, soft skills, conversational skills, leadership, etc.,

Human-in-control training. Automated training systems should integrate human-in-control training and emphasize the importance of human supervision in enhancing both learning outcomes and trust. Integrating features like periodic reviews by human experts can provide users with personalized feedback that aligns AI assessments with real-world expectations and addresses the trust gap. Importantly, human intervention should not be limited to expert supervision; users themselves should have the agency to contest AI assessments or modify system inputs [10, 350]. This agency fosters a sense of control and engagement in the training process. Additionally, designing AI systems to express uncertainty when encountering ambiguous situations can improve transparency and promote trust among users [34, 99]. This could include prompts that invite users for help after expressing uncertainties, thereby fostering a collaborative feedback loop. By acknowledging users' contributions, systems can improve AI performance and human well-being [204, 331]. This dual-layered involvement—both from experts and candidates—promotes a more dynamic and flexible training environment, ensuring that AI systems work as a supportive tool rather than a sole decision-maker.

Interactive and conversational training, like in the real world. Automated training systems should prioritise creating interactive, conversational experiences that mirror real-world dynamics. This design fosters engagement and ensures that users feel motivated to return for continued practice, review feedback, and improve their performance. For instance, Design A1 (Fig. 3.4) suggests a chat-like video conversation interface to conduct interviews. By incorporating features such as follow-up questions, context-aware responses, and real-time interaction, these tools can simulate the natural flow of an interview, making the experience more immersive and relatable. Integrating such conversational elements can enhance a user's sense of relatedness and belonging, which has been shown to influence intrinsic motivation to engage in training and skill development [266, 271].

Personalised and tailored training sessions with actionable feedback Automated training systems should focus on offering highly personalised and targeted training sessions tailored to specific job parameters being assessed. For example, these sessions can include the use of ideal example videos to help candidates understand the nuances of successful in-

terviews, as well as generating an *ideal* version of their responses, either as video or text, for comparison and improvement. Additionally, real-time cues on key areas like facial expressions, tone of voice, and body language can provide actionable, immediate feedback that allows users to adjust and refine their performance on the spot. This personalised approach not only improves learning outcomes but also ensures that each user receives feedback and training uniquely suited to their needs.

Responsible assessment by design Both job seekers and AI practitioners emphasised the importance of ethical, inclusive, and responsible assessment practices in automated interview training systems. As shown in our study, involving stakeholders early encourages fair and responsible design in automated training systems. Future designs should prioritise data diversity, algorithmic inclusivity, and privacy to ensure fair evaluations across different demographics, including age, race, gender, and cultural backgrounds, and regularly audit algorithms for potential biases. Additionally, systems should provide transparency and accountability in their decision-making processes, allowing users to understand *how* assessments are made and *why* certain decisions are taken. Ethical design must also include mechanisms for users to contest or challenge assessments, thereby promoting accountability and fostering trust in the system.

3.5.5 Limitations

Our study has some limitations. While we involve two key stakeholders—job seekers and AI practitioners—other important stakeholders, such as UX designers and recruiters, are not included. UX designers could add an additional layer of nuance by aligning the needs of job seekers with AI practitioners' technical expertise, ensuring a more seamless user experience. Additionally, expert job counsellors or recruiters could help define the parameters essential for interview training, offering valuable insights on the specific skills and behaviours to focus on. However, our methodology can be adapted to involve these additional stakeholders, and we encourage future research to do so, as this could lead to deeper insights and improved designs.

Another limitation is that we conducted separate sessions with job seekers and AI practitioners. It would be valuable to explore how results might differ or whether our findings hold if these stakeholders are brought together for joint discussions. Moreover, the study participants were predominantly young job seekers, with a lower representation of female AI practitioners. Different age groups and more diverse demographics could introduce new requirements and perspectives. Future studies could benefit from including a wider range of participants to better understand user needs and preferences.

3.6 Conclusion

In this paper, we employed a multi-phase co-design study with job seekers and AI practitioners, two key stakeholder groups in automated interview training systems. The co-design study aimed to understand job seekers' needs and align them with AI capabilities, incorporating insights from AI practitioners to identify challenges, desired functionalities, and differing perspectives in designing automated interview training systems. We contributed

to the design of automated interview training systems and broader design considerations for AI-assisted training tools. Our findings reveal alignment in a lot of perspectives and some fundamental tensions between the two groups. Future work should focus on alleviating some of these differences, such as advocating for accountability in the design of AI-assisted training tools. Our work calls for more nuanced participatory approaches involving multiple stakeholder groups for AI design.

Chapter 4

Invisible Filters: Cultural Bias in Hiring Evaluations Using Large Language Models

Abstract. Artificial Intelligence (AI) is increasingly used in hiring, with large language models (LLMs) having the potential to influence or even make hiring decisions. However, this raises pressing concerns about bias, fairness, and trust, particularly across diverse cultural contexts. Despite their growing role, few studies have systematically examined the potential biases in AI-driven hiring evaluation across cultures. In this study, we conduct a systematic analysis of how LLMs assess job interviews across cultural and identity dimensions. Using two datasets of interview transcripts, 100 from UK and 100 from Indian job seekers, we first examine cross-cultural differences in LLM-generated scores for hirability and related traits. Indian transcripts consistently receive lower scores than UK transcripts, even when anonymized, with disparities attributed to linguistic features such as sentence complexity and lexical diversity. We then perform controlled identity substitutions (varying names by gender, caste, and region) within the Indian dataset to test for name-based bias. These substitutions do not yield statistically significant effects, indicating that names alone, when isolated from other contextual signals, may not influence LLM evaluations. Our findings underscore the importance of evaluating both linguistic and social dimensions in LLM-driven evaluations and highlight the need for culturally sensitive design and accountability in AI-assisted hiring.

4.1 Introduction

Artificial intelligence (AI) has rapidly transformed the hiring landscape, reshaping how employers source, screen, and evaluate candidates [235]. A recent survey by Harvard Business School [114] highlights the widespread adoption of AI-enabled recruiting technologies: approximately 63% of employers reported using a Recruitment Management System (RMS), a figure that rises to 69% for organizations with over 1,000 employees. In the U.S., this

number is even higher at 75%, compared to 58% in the UK and 54% in Germany. These systems are not limited to administrative support. They are widely used to filter and rank candidates, with over 90% of employers employing them for initial screening.

Job interviews remain the most common selection method for employment [307], and digital platforms are being increasingly used for hiring [109]. Asynchronous video interviews (AVI), also called online video interviews [174], allow candidates to complete interviews at their convenience and are gaining popularity due to their scalability and ease of use. Many employers automate various aspects of the hiring process, including candidate scoring and interview scheduling. AVIs [256, 261] and video resumes [225] are used in early screening rounds[34], leveraging AI techniques [68] like computer vision, natural language processing, and audio processing to score candidates. As AI capabilities have expanded, large language models (LLMs) like GPT-4 [101] and Gemini [120] have emerged as particularly powerful tools in these pipelines. Their accessibility through APIs, user-friendly interfaces, and support for rapid prototyping [324] makes them especially appealing for companies that may have previously avoided AI tools due to cost or technical barriers. As a result, the integration of LLMs into hiring workflows is not only feasible but also likely inevitable.

Yet, as AI and LLMs in particular assume increasingly central roles in hiring decisions, there are urgent concerns about fairness, accountability, and cultural bias. Previous studies have demonstrated that text-to-image models [122, 250] and LLMs often encode and prioritize Western norms and values [4], leading to representational harms for non-Western users and contexts. These issues are especially pronounced in global South settings, such as South Asia, where Western-centric models can marginalize culturally specific forms of expression and communication [250]. Previous work has shown discriminatory hiring practices based on gender and race [68], including those using LLMs for resume screening tools [336]. While explicit forms of stereotyping in AI outputs are concerning, more subtle and insidious biases pose an equally serious threat. Recent research shows that LLMs can shape user behaviour through autocomplete and writing suggestions, altering not just syntax but also tone, content, and even attitudes toward social and cultural issues [4, 141, 247]. This raises a critical, yet underexplored question: do such cultural biases extend to how LLMs evaluate texts in high-risk contexts, such as job interviews?

The field of fairness and ethics has increasingly called for a rethinking of how AI harms are conceptualized and studied. Scholars have emphasized the need for community-centered approaches and for frameworks that move beyond Western-centric notions of fairness. However, this remains an under-addressed issue, as research into ethics and fairness often centers Western perspectives and relies on Western frameworks of ethics and fairness [222, 284]. A shift toward more inclusive and context-sensitive evaluation is crucial if we are to understand how AI systems behave across diverse cultural and social contexts, particularly in high-stakes domains such as employment.

Building on this need for globally grounded evaluation frameworks, we examine whether cultural biases extend to LLM-based assessments of job interviews, by comparing interview responses from two culturally distinct English-speaking contexts: the United Kingdom and India. This pairing enables us to study LLM behaviour across Western and South Asian settings, which differ in their social and cultural norms and are often unevenly represented

[34]For example, `www.hirevue.com`, `www.modernhire.com`, `www.viasto.de`. Last accessed August 2025.

[284]. The UK is frequently associated with more individualistic orientations, often prioritizing efficiency and achievement. India reflects more collectivistic tendencies, placing greater emphasis on relationality and kinship [139, 140, 306]. These distinctions provide a meaningful contrast for understanding whether LLMs systematically favor patterns associated with Western contexts. By focusing on the UK and India, the study aims to investigate how AI systems behave across culturally diverse populations and whether Western-centric biases influence automated assessments in hiring.

This paper makes three key contributions to understanding cross-cultural variation and potential bias in LLM-based evaluations of job interview transcripts.

1. *Cross-Cultural Comparison of LLM Evaluations:* We analyze two corpora of interview transcripts—100 from UK-based job seekers and 100 from Indian job seekers—to examine how LLMs assess responses across Western and non-Western contexts. By anonymizing identity-related entities, we isolate linguistic and semantic features as the primary source of variation. We evaluate each transcript on job attributes like hireability, positive impression, self-promotion, and use of stories, considered important in an interview, using two widely popular LLMs, GPT-4o and Gemini Flash1.5. Results show that Indian transcripts consistently receive lower scores than UK transcripts across all evaluated dimensions, raising concerns about systematic scoring disparities.

2. *Linguistic Feature Analysis:* To understand the basis of these scoring disparities, we extract a range of linguistic features like type-to-token ratio, sentence length, confidence markers (e.g., hedging vs assertive language), and examine their correlation with LLM-generated scores. Our findings indicate that lexical diversity, syntactic complexity, and readability strongly influence evaluations. Importantly, even after controlling for these features, UK transcripts continue to score higher, suggesting culturally conditioned preferences embedded in LLM behaviour.

3. *Controlled Identity Substitution to Probe Bias*: Previous work has shown that popular LLMs show bias towards marginalized Indian groups [162]. We further explore whether identity-based cues influence LLM evaluations by systematically altering names within the Indian dataset to signal gender, caste, and regional identity, while holding all other content constant. These controlled substitutions reveal no statistically significant differences in LLM scores, suggesting that names alone, in the absence of additional contextual cues, may not be sufficient to trigger bias. These findings offer a cautiously optimistic signal in the ongoing discussions on fairness and highlight the utility of controlled counterfactual designs for bias testing.

It is important to clarify that this study is not intended as a comprehensive evaluation of LLM capabilities in hiring contexts. We do not present our findings as definitive assessments of any specific model or of LLMs in general. Rather, this targeted investigation serves as an illustrative exploration of how identity cues and linguistic differences across cultures may influence AI-generated evaluations. Despite its limitations, the study highlights critical considerations for the design and deployment of LLMs in high-stakes domains like hiring and invites deeper reflection on fairness, transparency, and the risks of AI in critical decision-making.

4.2 Related Work

We review scholarly work examining cultural bias in AI systems, particularly how large language models (LLMs) prioritize Western linguistic norms, values, and communication styles. Given our focus on AI-assisted evaluation of written job interview responses, we also review literature on the evaluation of hiring technologies powered by LLMs. This intersection of sociotechnical systems, language, and cross-cultural evaluation frames the core concerns of our work: how seemingly neutral AI recommendations may carry invisible cultural filters that influence high-stakes decision-making processes such as hiring.

4.2.1 Cultural Bias in LLMs

Large language models are known to encode cultural and linguistic biases present in their training data. Because most LLMs are trained on English corpora dominated by Western sources, they tend to favour Western norms and values in their outputs [17, 155], primarily when prompted in English [58]. For instance, early analyses revealed that GPT-3 exhibited severe religious and ethnic biases, such as disproportionately associating Muslims with violence [2].

Recent studies using cross-cultural benchmarks have confirmed systemic Western alignment in LLM outputs. Tao et al. [304] showed that state-of-the-art LLMs align more closely with Western values (as measured by World Values Survey data) than non-Western ones. Naous and Xu [221] attribute these patterns to pre-training data imbalances and tokenization mismatches, noting that even multilingual LLMs favour Western-associated entities when operating across languages. Khandelwal et al. [162] show that when prompted with Indian context data, popular LLMs exhibit a strong propensity to produce stereotyped outputs about marginalized Indian groups (e.g., caste and religion), far more than the bias observed on traditionally studied Western axes like gender or race.

LLMs also tend to homogenize culturally distinct expressions. Agarwal et al. [4] found that AI writing assistants encouraged Indian English speakers to adopt U.S.-centric phrasings, diminishing local nuance. Similarly, Mukherjee et al. [217] expanded bias audits to 24 languages, revealing that social biases are not globally uniform: bias in models can manifest differently in, say, Indian languages versus European languages, and along dimensions like toxicity or ableism that extend beyond the gender/race categories typically examined.

Outside text, cultural erasure extends to visual generative models. Qadri et al. [250] found that text-to-image models presented South Asian subjects through a Western, stereotypical gaze. Ghosh et al. [122] echoed this finding, showing generative AI produced exoticized and inauthentic visuals when rendering non-Western prompts. These problems are particularly evident in global South contexts, where Western-centric frameworks often marginalize culturally specific forms of expression and communication.

Researchers have also explored bias tied to dialect and intersectionality. Deas et al. [81] evaluated model performance on African American English (AAL), finding significant drops in generation quality compared to standard English. Sandoval et al. [278] noted that Black users felt current LLMs failed to capture the nuance of AAL and preferred systems capable of respectful register-shifting. Ma et al. [197] further demonstrated that LLMs reproduce

intersectional stereotypes (e.g., combining religion, race, and gender) despite safety training.

These nuanced results show that LLMs can reproduce societal stereotypes targeting complex, intersecting identities, not just single demographics. Prior research establishes a growing awareness that cultural and linguistic biases are *baked into* LLM behaviour. Our work extends this literature by examining such biases in an AI-driven hiring evaluation setting to examine the influence of cross-cultural subtleties (like a candidate's style or cultural background) in LLM judgments.

4.2.2 Bias in Hiring with AI

Algorithmic hiring systems have repeatedly been shown to replicate existing human and institutional biases. Raghavan et al. [252] reviewed commercial hiring algorithms and found that bias-mitigation claims were often unsubstantiated, as vendors frequently failed to disclose meaningful technical documentation, offered no empirical evidence of fairness improvements, and provided vague or inconsistent definitions of bias. Their analysis revealed that many systems marketed as *fair* relied on proprietary scoring mechanisms and opaque inputs, making it difficult for external auditors or employers to assess the legitimacy of fairness claims. Chen [68] similarly observed that many recruitment platforms inadvertently inherit biases from training data and developer assumptions. Hangartner et al. [130] developed a cost-efficient methodology to monitor hiring discrimination on online recruitment platforms and found that application callbacks are lower for immigrant and minority ethnic groups.

Numerous real-world incidents have documented the risks of algorithmic bias in hiring: from reinforcing historical inequalities[35] to employee terminations based on automated assessments[36] [207]. One infamous example is Amazon's discontinued hiring model, which penalized resumes containing terms like "women's chess club" due to training on male-dominated hiring data.[37] Traditional correspondence tests, where researchers send fictitious but identical CVs except for the minority trait to be tested (for example, names that are deemed to sound 'Black' versus those deemed to sound 'white'), have quantified discrimination in hiring practices [31]. Extending these concerns to AI systems, Wilson and Caliskan [336] found that LLM-based resume screening disadvantaged Black and female-associated names, even when all other resume content was identical. Similarly, Puutio and Lin [249] ran 2,000 simulated applications through ChatGPT and found emergent patterns of discrimination, such as preference for elite education or positional bias in response order.

To address these issues, researchers have proposed a range of mitigation strategies. Specific to bias in hiring Albaroudi et al. [8] present a taxonomy of fairness techniques, including de-biasing embeddings and preprocessing structured data. Fabris et al. [105] advocate for a multidisciplinary approach that incorporates legal and ethical oversight alongside technical solutions. More broadly, the literature outlines strategies ranging from practical interventions to reduce societal harms [75, 171] to frameworks for identifying and documenting risks and biases [86, 258, 324], with the goal of improving practitioner awareness and guiding responsible system design.

[35] https://incidentdatabase.ai/cite/135/

[36] https://incidentdatabase.ai/cite/111/

[37] https://incidentdatabase.ai/cite/37/

Table 4.1: **Metrics used for LLM-based hiring evaluation:** Each metric, its description, and its source from literature are shown. This, with corresponding items, was provided to the LLM in the prompt alongside the interview transcript.

Metrics	Source questionnaire	Items (N)	Description
Hireability	Madera et al. [202]	3	Likelihood of an individual getting hired, measures the potential that a job seeker will be an effective employee [202].
Positive impression	Rao et al. [257]	3	Lasting impact an individual has after the interview, influenced by non-verbal behavior [151].
Self-promotion	Bolino and Turnley [41]	4	Impression management tactics; refers to an attempt to evoke favourable character perceptions about oneself to elicit respect for one's abilities rather than just being liked [123, 326].
Storytelling	Rao et al. [257]	1	Answering past behaviour questions with a narrative, providing information on the situation, task, action and results [21].

Research Gap

While prior research on AI hiring bias has mostly focused on race and gender within a single-cultural context (often the U.S.), we extend the inquiry to cross-cultural disparities between Indian and UK candidates. This study builds on the cultural bias findings from the NLP domain and the fairness concerns in algorithmic hiring. By bringing these threads together, we aim to shed light on whether candidates from non-Western or non-mainstream backgrounds face an invisible disadvantage when assessed by LLMs. In summary, our related work review indicates that (1) LLMs carry latent cultural biases favoring specific values, dialects, and demographic stereotypes, and (2) uncritical use of AI in hiring has led to measurable discrimination. These two insights directly motivate our study: how LLM-generated hiring evaluations might inadvertently prefer some cultural-linguistic style over others, and in doing so, our findings contribute to the broader effort of ensuring fairness and cultural equity in AI-supported hiring practices.

4.3 Method

To examine how LLMs evaluate job interview performance across cultural and identity contexts, we collected and transcribed video interviews from Indian and UK job seekers. The resulting transcripts were analyzed under two contexts: (i) anonymized evaluations and (ii) identity-modified evaluations.

4.3.1 Participant Recruitment and Data Collection

A priori power analysis indicated that a minimum of 143 participants was required to detect a medium effect size ($f = 0.40$) with $\alpha = 0.05$ and $power = 0.95$. To account for potential

technical issues during recording, we recruited a total of 200 participants. Screening criteria included: (i) actively seeking employment; (ii) unemployed, partly employed, or not in paid work; (iii) at least 18 years of age; (iv) willingness to record a video interview; (v) access to a webcam.

4.3.1.1. Indian Data We recruited 100 participants from our university, of whom 42 were female and 58 were male. All participants were Indian nationals in the age group of 18-23. All participants were students and reported professional fluency in English.

4.3.1.2. UK Data We recruited 100 participants via the Prolific platform[38]. We restricted the participants to those who resided in the United Kingdom and Ireland, with English as their first language. The sample included 64 females and 36 males. Age distribution was: 18–29 ($n = 50$), 30–39 ($n = 31$), 40–49 ($n = 14$), 50–59 ($n = 3$), and 60+ ($n = 2$). Their nationalities included the UK ($n = 93$), Ireland ($n = 4$), and the US ($n = 3$). 64 participants were students.

4.3.1.3. Procedure Participants completed asynchronous video interviews using a custom web-based platform. Each session included five questions: one introductory prompt (e.g., "Tell me about yourself") and four additional questions assessing the interviewee's competence and knowledge. These questions were randomly drawn from a pool spanning six commonly used structured interview domains: work experience, leadership, personality, career goals, skills, and behavioural/situational responses [131, 334]. We simulated a first-round behavioral job interview for a generic professional role, where no specific job title or responsibilities were included, to mirror common HR screening practices and allow for consistent, role-agnostic evaluation of candidate responses. Each response was limited to a maximum of five minutes. All videos were transcribed using the OpenAI Whisper API base model [251].

4.3.2 LLM-Based Hiring Evaluations

4.3.2.1. Metrics We assessed candidate performance on four job-relevant attributes identified in personnel selection research: *hireability*, *positive impression*, *self-promotion*, and *storytelling*. These attributes were selected for their relevance to hiring decisions based on personnel selection literature [147] and for being suitable for text-based evaluation. We used standardized scales to measure each of them, with each attribute having multiple item descriptors scored on a five-point Likert scale (see Table 4.1 for details).

4.3.2.2. LLM Scoring All transcripts were anonymized by replacing identifiable entities (e.g., names, cities) with generic placeholders (e.g., Person X, City Y). We crafted and iteratively refined a scoring prompt instructing the LLM to rate each transcript against four job-relevant metrics. The prompt included the definitions of the metrics, corresponding items

[38] `https://www.prolific.co/`

(see Table 4.2), and their respective scale from 1 to 5. The LLM was tasked with embodying the role of a recruiter, making hiring decisions without speculating on the candidate's identity or background. Like the data collection process, the prompt simulated a first-round behavioral interview for a generic role, intended to screen candidates for subsequent, job-specific interview stages. Each transcript was independently scored by two LLMs: OpenAI's GPT-4o and Google's Gemini Flash 1.5. Final scores for each attribute were computed as the mean of its corresponding items.

4.3.2.3. Linguistic Features To investigate whether linguistic characteristics contributed to observed cross-cultural differences in LLM scores, we extracted several standard textual features known to influence text-based assessments.

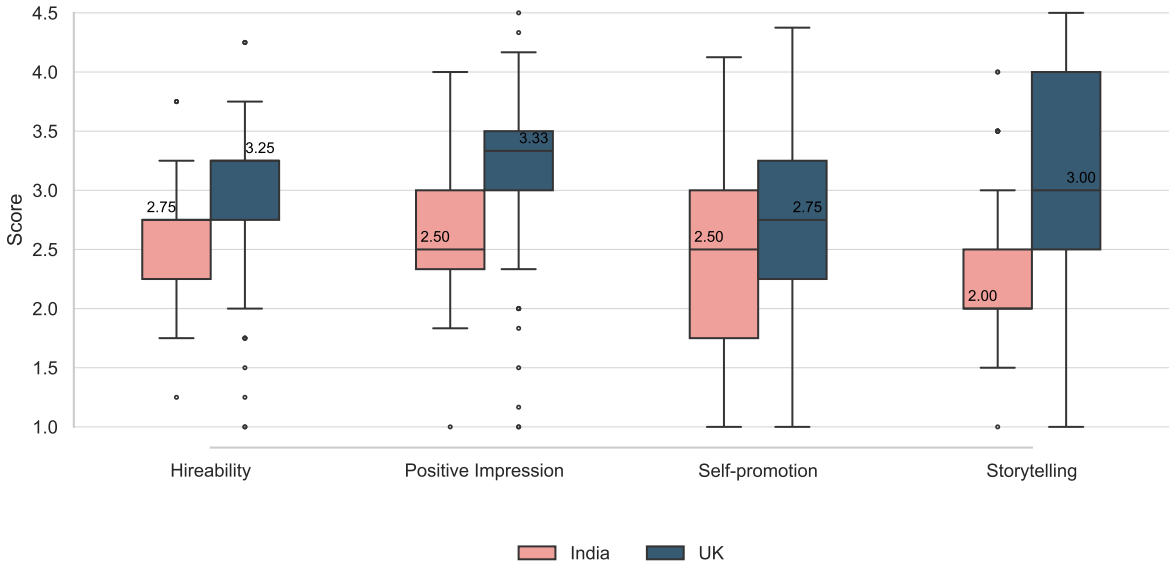


Figure 4.1: Box plots showing scores for four job-relevant metrics as scored by GPT-4o. LLM scores are significantly lower for Indian transcripts than the UK across all four metrics, indicating a potential bias favouring Western linguistic styles.

1. *Lexical diversity* was measured using the type-token ratio (TTR), defined as the number of unique words divided by the total number of words, indicating vocabulary diversity.

2. *Sentence complexity* was operationalized using average sentence length (ASL), computed as total word count divided by the number of sentences.

3. *Readability* was quantified using the Flesch Reading Ease (FRE) score [111], calculated as:

$$\text{FRE} = 206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW})$$

where ASL is the average sentence length, and ASW is the average number of syllables per word. Higher scores indicate easier readability.

4. *Confidence markers* were quantified by calculating the frequency of hedging terms (e.g., "maybe," "perhaps") and assertive terms (e.g., "clearly," "certainly"), expressed as ratios relative to total word count. These measures reflect the speaker's projected level of confidence in their responses [150].

4.3.2.4. Identity-Based Substitutions To assess the influence of identity cues, we performed controlled substitutions of candidate names within the Indian dataset. Using a within-subjects design, we generated identity-modified versions of each transcript by varying candidate names along three dimensions: gender (male/female), region (North/South Indian), and caste (upper/lower), informed by culturally salient naming conventions [100]. We chose 5-10 names per combination (e.g., 10 names for a North Indian upper caste female), leading to a total of 60 names. For each transcript, we substituted a randomly selected name from each combination. All other content remained identical, and anonymized entities, except for names, were unchanged. This produced multiple identity variants per transcript (e.g., the same content with a North Indian upper-caste male name vs. a South Indian lower-caste female name). To reduce environmental impact and computational cost, we randomly selected 50 transcripts from the Indian dataset, resulting in 400 data points (50 transcripts $\times$ 8 identity conditions), and evaluated only using GPT-4o. This enabled the evaluation of LLM sensitivity to socio-cultural identity signals embedded in names.

4.3.2.5. Clarifying the Cross-Cultural Comparison This study compares LLM-generated evaluations of anonymized interview transcripts from India and the United Kingdom, focusing specifically on differences in how LLMs score job-relevant attributes and how linguistic features vary across the two datasets. As the study uses text-only transcripts, the comparison is restricted to verbal content; other modalities, such as accent, prosody, and non-verbal behavior, were not analyzed, although they may also influence interview performance.

The two samples differ in demographic composition: all Indian participants were students aged 18–23, whereas the UK sample spans a broader age range and includes both students and professionals. These differences may contribute to variation in communication style and linguistic complexity, and they are acknowledged as potential confounds that may also explain part of the observed disparities.

We used widely deployed general-purpose LLMs (GPT-4o and Gemini Flash 1.5), which reflect models currently used in commercial hiring systems. Culturally adapted or region-specific LLMs were not tested. It remains an open question whether culturally aware models would mitigate or reproduce such disparities, and this represents an important direction for future work.

4.3.2.6. Analysis To evaluate cross-cultural differences in LLM-generated scores, we compared the outputs for the anonymized Indian and UK transcripts. Due to technical issues in some recordings, we analyzed 99 Indian and 95 UK transcripts. We used the Shapiro–Wilk test to confirm that the score distributions deviated from normality. Consequently, we applied the non-parametric Mann–Whitney U test to assess differences in LLM-assigned scores between the two groups. This analysis was conducted independently for each LLM (GPT-4o and Gemini).

Table 4.2: **Statistical significance of the results for cross-cultural comparison:** Mann-Whitney U test results between the Indian and UK LLM-generated scores. The test showed statistically significant differences across all the chosen job-relevant metrics with GPT-4o, except for self-promotion with Gemini Flash 1.5, indicating scoring disparities favouring Western-style responses. Overall final scores for each metric were computed as the mean of its corresponding items. $^{**}p<0.01$; $^{***}p<0.001$; ns:$p > 0.05$.

Metric	**GPT-4o**		**Gemini Flash 1.5**	
	statistic	p-value	statistic	p-value
Overall Hireability	2284.5	***	1888.5	***
As a recruiter, I would be willing to hire him/her? (H1)	2314.0	***	2005.5	***
As a recruiter, I would perceive him/her as a "top-notch" candidate. (H2)	2261.0	***	1932.0	***
As a recruiter, I would think he/she would make an effective employee. (H3)	2295.0	***	1993.5	***
As a recruiter, I would consider him/her as an "excellent" asset. (H4)	2280.0	***	2096.0	***
Overall Positive Impression	2168.5	***	1913.5	***
The performance during the interview was good. (P1)	2268.5	***	2015.5	***
The interviewee gave a good impression of him/herself. (P2)	2233.5	***	1848.5	***
The interviewee looked competent. (P3)	2316.5	***	2585.5	***
Overall Self-Promotion	3646.5	**	4406.0	ns
Talk proudly about their experience or education. (S1)	4078.5	ns	5000.0	ns
Make people aware of their talents or qualifications. (S2)	3619.5	**	4389.5	ns
Let others know that their are valuable to the organization. (S3)	3329.0	***	4036.0	ns
Make people aware of their accomplishments. (S4)	3582.5	**	4220.0	ns
Storytelling The interviewee provided a clear and concise story. (ST1)	1974.0	***	2346.5	***

To further investigate whether linguistic properties contributed to the observed scoring disparities, we conducted a post hoc multiple linear regression analysis using the combined dataset of Indian and UK transcripts. The dependent variables were the four job-relevant metrics. The predictors included region (India vs. UK) along with linguistic features: type-token ratio, average sentence length, Flesch Reading Ease (FRE), hedging ratio, and assertive ratio. This model enabled us to determine whether region remained a significant predictor of scores after controlling for textual features. Prior to fitting the model, we checked standard regression assumptions. Residual diagnostics indicated mild non-linearity and deviation from normality, confirmed by a significant Shapiro-Wilk test ($p = 0.0003$). Multicollinearity was assessed using VIFs; most predictors were well below conventional thresholds, though average sentence length and FRE showed high correlation ($r = -0.93$). Both were retained, given their relevance and contribution to model fit.

To examine potential identity-based bias within the Indian dataset, we analyzed score changes across identity-substituted variants. A separate linear mixed-effect regression (LMER) model was fitted for each of the four job-relevant metrics with gender, region, and caste (based on name cues) as fixed effects and transcript ID as a random effect to account for repeated measures across the eight identity conditions. This allowed us to test whether socio-cultural identity signals embedded in names influenced LLM evaluations, independent of content.

4.4 Results

4.4.1 Cross-Cultural Differences in LLM Evaluations

We compared the anonymized interview transcripts from Indian and UK participants to assess whether LLMs exhibit differential scoring across cultural contexts. Mann–Whitney U tests were conducted separately for GPT-4o and Gemini Flash 1.5 on each of the four job-relevant metrics—hireability, positive impression, self-promotion, and storytelling—and their component items.

For both LLMs, Indian transcripts consistently received significantly lower scores (Figure 4.1) than UK transcripts on hireability, positive impression, and storytelling, with all p-values well below 0.001, indicating strong evidence of a scoring gap. Refer to Table 4.2 for details. Notably, all the corresponding items (H1-H4, P1-P3, ST1) for the three metrics were scored consistently lower. The median scores ranged between 2 to 2.75 for Indian transcripts, whereas for UK transcripts, they ranged between 2.75 to 3.25, indicating a large gap. These findings suggest a potential structural bias in LLM evaluations favouring UK linguistic norms or response styles. In the case of self-promotion, GPT-4o showed significant differences for the overall metric and most sub-items (S2, S3, S4), while Gemini Flash did not show statistically significant differences on any self-promotion measure.

4.4.2 Influence of Linguistic Features

To assess whether linguistic properties explained variation in LLM-generated evaluations, we fitted multiple linear regression models on combined data predicting each job-relevant

Table 4.3: **Regression models for linguistic feature analysis**: Coefficients of the regression models built with linguistic features predicting the job-relevant metrics: hireablity, positive impression, self-promotion and storytelling. Each predictor has its coefficient on the first line and the t-statistic on the line below. The last row indicates Adjusted R^2 and F-statistics of the model with degrees of freedom. *p<0.05; **p<0.01; ***p<0.001

Predictor	**Hireability**	**Positive Impression**	**Self-Promotion**	**Storytelling**
Intercept	4.218***	4.300***	4.992***	3.860***
	t = 9.58	t = 8.99	t = 8.74	t = 6.60
Region [UK]	0.444***	0.490***	0.275**	0.826***
	t = 5.36	t = 5.45	t = 2.56	t = 7.51
Type-token ratio	-0.656	-0.415	-1.753***	-1.227*
	t = -1.73	t = -1.01	t = -3.57	t = -2.44
Avg. sentence length	-0.014**	-0.016**	-0.019**	-0.012*
	t = -3.13	t = -3.26	t = -3.31	t = -2.00
Flesch Reading Ease	-0.016***	-0.017***	-0.022***	-0.013*
	t = -3.83	t = -3.65	t = -3.92	t = -2.35
Hedging ratio	0.127	0.140	0.080	0.176
	t = 1.24	t = 1.27	t = 0.61	t = 1.30
Assertive ratio	0.196	0.208	0.245	0.267
	t = 1.81	t = 1.77	t = 1.75	t = 1.86
Adjusted R^2	0.233	0.221	0.170	0.294
F Statistic (df = 6; 187)	10.77***	10.11***	7.59***	14.40***

metric from region (India vs. UK) and a set of linguistic features: type-token ratio, average sentence length, Flesch Reading Ease, hedging ratio, and assertive ratio.

The results of the regression analysis can be found in Table 4.3. The regression models for hireablity, positive impression, self-promotion, and storytelling were all statistically significant. Across all models, region consistently emerged as a significant predictor, with UK transcripts receiving higher scores than Indian transcripts, even after controlling for linguistic characteristics. The scores were > 0.4 points higher for hireablity, positive impression, and storytelling compared to Indian transcripts.

Among the linguistic predictors, both average sentence length and FRE, i.e., easier-to-read language, were consistently and negatively associated with all LLM scores. While these two features are partially correlated, since longer sentences reduce reading ease, they each contribute uniquely to model predictions. The negative coefficients suggest that transcripts with simpler, more readable language (i.e., higher Flesch scores) were rated less favorably, while those with longer sentences were also penalized. In other words, LLMs may reward sophisticated vocabulary use (with lower Flesch scores) while penalizing excessive sentence length. This may penalize speakers from linguistic backgrounds where longer, more formal sentence constructions are common (e.g., Indian academic English), thus contributing to unintended cultural bias [28, 110, 169].

Interestingly, the type-token ratio (a proxy for lexical diversity) was also significantly and negatively associated with self-promotion and storytelling. When clarity in articulating accomplishments and narrative coherence is crucial, excessive lexical diversity may not be favoured. Rather than rewarding lexical variety alone, LLMs may favor responses that maintain coherence and relevance while using moderately sophisticated vocabulary. Taken together, these findings point to a nuanced preference for concise, semantically dense language, rather than for either simplicity or complexity in isolation.

4.4.3 Effect of Identity Cues

We tested whether identity cues of gender, region, and caste embedded in candidate names affected LLM-generated evaluations of Indian interview transcripts. Linear mixed-effects models were fitted for each job-relevant metric across 400 data points, with transcript ID included as a random effect. Across all four models, neither region nor caste had a statistically significant effect on scores ($p > .10$), indicating minimal sensitivity to these identity dimensions. Gender showed small but statistically significant effects for hireability ($p = .031$) and positive impression ($p = .034$). However, these differences were not significant in post-hoc pairwise comparisons with TukeyHSD [308] (e.g., $p = .275$ for hireability), suggesting the effects may not be robust. Gender effects were non-significant for self-promotion ($p = .149$) and storytelling ($p = .095$).

Overall, these findings suggest that LLMs did not consistently alter their evaluations based on identity cues in names alone. These findings suggest that LLMs, when stripped of broader socio-contextual inputs, may not rely heavily on stereotypical associations tied to names alone in controlled textual settings.

4.5 Discussion

This study examined how large language models (LLMs) evaluate job interview performance across cultural and identity contexts, using systematically controlled interview transcript data from Indian and UK job seekers. Our analysis yielded three core findings: (1) a consistent scoring disparity disadvantaging Indian transcripts, (2) systematic associations between linguistic features and LLM evaluations, and (3) minimal effects of identity-based name cues in controlled settings. Below, we discuss these findings and their broader implications.

4.5.1 Systematic Cultural Disparities and Consequences for Workforce

Our cross-cultural comparison revealed that Indian transcripts received significantly lower scores than UK transcripts. These differences persisted across two popular LLMs and were robust to anonymization of identity-related content. Our findings suggest that LLM-based interview scoring can inadvertently favour Western linguistic patterns and communication norms (e.g., phrasing, tone), leading to systematic disadvantages for non-Western candidates. The immediate concern is that such biases could translate into less diverse hiring outcomes. If organizations rely on these AI evaluations for early-stage screening, qualified candidates who don't fit the model's learned linguistic mold may be filtered out, perpetuating the underrepresentation of non-Western groups in global hiring pipelines. Our findings expand on the previous studies showing favourable outcomes for Western contexts [336] where an LLM-based resume screener significantly favoured White-associated names while consistently disadvantaging Black male candidates. In effect, "hidden" talent from marginalized communities can remain hidden [114] if AI tools systematically overlook or undervalue non-normative styles and experiences. At scale, these exclusions reinforce global inequalities, risk legal and reputational harm for organizations, and replicate historical patterns of discrimination. These structural biases not only raise ethical and legal red flags but also threaten to create a workforce that is homogeneous and missing perspectives from broad segments of society.

Furthermore, if these LLMs are used not only for evaluation but also for designing interviews [254] or supporting interview training and preparation [69, 257], they risk introducing an additional layer of bias reinforcement. AI-driven interview coaching tools, often trained on Western-centric data, may unintentionally pressure candidates to adopt Western communication styles, risking the loss of culturally rooted expression. While such tools can support skill development, they may penalize authenticity: candidates must either conform to dominant norms to improve their scores or maintain their natural style and risk lower evaluations. Our results hinted at this effect: Indian candidates might need to modulate their natural speech patterns to score well under the LLM evaluator's criteria. More direct evidence comes from related studies of AI writing assistants. For example, Agarwal et al. [4] found that when a Western-centric LLM provided suggestions, Indian participants gradually adopted more Westernized writing styles, losing some of the linguistic nuances that reflected their cultural identity. Over time, such homogenization may suppress diverse ways of thinking and communicating in professional contexts. This highlights the need for AI coaching

systems that value varied expressions of eloquence and competence, rather than reinforcing a singular cultural norm.

4.5.2 AI Trust and Overreliance in Global Contexts

Sociopsychological studies have found that individuals in more collectivist cultures often view AI as a helpful, authoritative extension of their community or self, whereas those in individualistic cultures are more likely to treat AI with skepticism, seeing it as a potential threat to personal autonomy [23]. In practice, this means an HR manager in India or another collectivist context might place higher faith in an AI's interview ratings, accepting them with little question, while a UK or US manager might be more inclined to double-check the machine's judgments. If users assume the AI is neutral and objective, they may overlook subtle cultural prejudices influencing the scores. Overreliance is particularly dangerous when the model's biases align with the user's blind spots: for example, a Western-trained model and a non-Western recruiter both accepting Western norms as default.

Introducing explainability into the AI's evaluations offers one promising route to mitigate this risk. Research on human-AI decision-making has shown that transparent explanations can reduce overreliance on AI outputs by encouraging users to critically evaluate recommendations [316]. In the hiring context, if an LLM explains the reasoning behind a low score, such as pointing to specific communication traits or responses, a culturally aware recruiter can assess whether the feedback reflects a genuine job-relevant issue or a cultural mismatch. This transparency empowers human stakeholders to interrogate and override the AI's judgment based on local norms and values.

More broadly, we argue that AI systems for hiring should be designed with intentional friction [209]. While conventional UX principles emphasize seamless integration, reducing user effort, such smoothness can obscure the limitations and cultural assumptions embedded in AI evaluations. Instead, we advocate for a *seamful design* [99] approach, which reveals the system's boundaries, uncertainties, and imperfections to users. Rather than concealing these seams, designers can use them to prompt critical reflection and contextual adaptation. While Explainable AI has traditionally focused on illustrating algorithmic logic, seamful design can extend this by revealing sociotechnical and cultural mismatches, encouraging users to question not just how the AI works, but whether it aligns with their context and values.

4.5.3 Toward Culturally Aware LLM Development

Mitigating cultural harms requires rethinking how we develop and deploy LLMs in hiring. Most AI bias and fairness evaluations to date have been Western-centric [276, 284, 311], often overlooking biases that matter in other regions. Khandelwal et al. [162] introduced the Indian-BhED dataset to evaluate caste and religion-based stereotypes, dimensions largely absent from standard fairness benchmarks. Their analysis revealed that popular LLMs frequently reproduced harmful stereotypes related to caste and religion, even more so than for attributes like gender or race that are commonly studied in the West.

In contrast, our study did not observe significant differences in LLM evaluations with

identity substitutions based on caste and geographic region within the Indian context. This discrepancy may stem from the limited ability of LLMs to detect subtle cultural signals embedded in names alone, or from the highly controlled nature of our data, where anonymization and content uniformity may have masked stereotype effects. These findings underscore the importance of developing and applying localized benchmarks and evaluation frameworks that can capture culturally specific biases, embedding different contexts (beyond names) of underrepresented or marginalized populations.

Equally important is a community-centered evaluation strategy when assessing and improving model behavior. Rather than AI developers defining what constitutes "fair" or "acceptable" output, recent research demonstrates the value of involving the communities that the AI impacts [122, 250]. This can lead to a deeper understanding of AI's blind spots and more culturally aware interventions.

Achieving equitable AI-driven hiring requires participatory approaches to LLM development that center on global and culturally diverse perspectives. For developers of LLM-based hiring tools, this can include expanding training data to reflect varied dialects, communication styles, and languages, and co-creating evaluation criteria with input from non-Western recruiters and job seekers. Model performance should be routinely audited using region-specific benchmarks (e.g., Indian-BhED) and informed by feedback from target user communities. For HCI and AI ethics researchers, this involves engaging a broad set of stakeholders like job seekers, employers, resource groups, social scientists, and policymakers throughout the AI lifecycle to ensure community-centered design and evaluation. On the regulatory front, standard bodies and policymakers can mandate cross-cultural bias impact assessments [38] for high-stakes AI systems. For instance, organizations could be required to audit AI interview platforms for disparate impact across cultural groups and provide transparent documentation on how language variation is handled. Requiring explainability in AI decisions, as discussed earlier, could be part of these standards to expose culturally misaligned reasoning.

Ultimately, fostering culturally aware LLMs demands an expanded view of fairness, one that incorporates global voices and lived experiences [276]. By involving diverse communities in shaping AI systems, we not only improve fairness and accuracy but also affirm the agency and satisfaction of those navigating AI-mediated hiring.

4.5.4 Limitations and Future Work

While this study provides important insights into the cultural dynamics of LLM-based hiring evaluations, some limitations warrant consideration.

ASR Limitations. The accuracy of interview transcripts was constrained by the use of an automatic speech recognition system (OpenAI Whisper), which can exhibit reduced accuracy for non-Western accents. A manual review of 30% of transcripts showed most errors involved named entities rather than substantive content. As a corrective measure, we anonymized all named entities prior to analysis. However, subtle transcription errors may still have affected linguistic features, particularly in the Indian dataset. Using accent-aware ASR models or fully human-corrected transcripts could reduce these effects.

Participant and Data Limitations. The Indian and UK participant groups differed de-

mographically: all Indian participants were university students aged 18–23, while the UK group spanned a broader age range, with half over 30 and potentially more professional experience. Though all participants from both groups were job seeking and preparing with self-paced interview preparations, this could have influenced their performance. Future studies should aim to match participants more closely on age or professional background to better isolate cultural effects from demographic influences.

To preserve privacy, all identifiable entities (other than names) universities, cities, and organizations, were anonymized. While necessary, this reduced context may have reduced the realism and richness of candidate responses, especially in sections where institutional affiliations or geographic ties contribute to stereotypes in self-presentation.

LLM shortcomings. A potential source of bias lies in the prompt design itself. To minimize subjective framing, we structured the task around granular, literature-derived items for each metric. Nevertheless, prompt design decisions may still embed implicit author biases. Future studies could mitigate this through prompt co-design with domain experts or prompt randomization across LLM runs.

Finally, our results reflect the behavior of two specific LLMs (GPT-4o and Gemini Flash 1.5). As model architectures, training data, and fine-tuning strategies evolve rapidly, future research should test whether these patterns generalize across other models and under fine-tuning with culturally diverse and demographically balanced datasets.

Scope of the Study. We focused on generic behavioral interviews rather than technical or role-specific assessments to model early hiring stages where general attributes such as communication and self-presentation are most relevant. Future studies should extend this framework to later stages to examine whether cultural biases persist in specialized contexts.

We did not establish a human ground-truth baseline, making it possible that some score differences reflect genuine performance variation. Comparative evaluations by HR professionals across cultural contexts could calibrate LLM scores and quantify any penalties relative to human assessments. Nonetheless, given that both groups were randomly selected job seekers, the UK group's consistently higher scores across all metrics remain concerning.

The study relied solely on transcribed text, excluding multimodal cues such as tone or body language that could amplify cultural differences in storytelling or self-promotion. Future work should incorporate multimodal inputs and embed identity cues beyond names.

Finally, our findings stem from a controlled evaluation setting; real-world hiring introduces further complexities, including recruiter persona, multi-stage processes, and AI–human interaction effects. Future work should assess how LLM scoring influences actual hiring outcomes and whether interventions such as explainable feedback improve fairness and trust.

4.6 Adverse Impact Statement

This study highlights the potential for large language models (LLMs) to systematically disadvantage candidates from non-Western cultural backgrounds in automated hiring evaluations. While our findings raise concerns about fairness, they are intended to inform the responsible development and use of LLMs in evaluative settings. Our experimental setup involved controlled comparisons and anonymized data to minimize harm to participants. Nevertheless,

we acknowledge that releasing findings about bias could be misused to justify exclusionary hiring practices or to downplay the significance of other forms of bias not observed in our controlled settings. We strongly discourage such interpretations and emphasize that the goal of this work is to improve transparency, equity, and accountability in AI-driven hiring systems. We recommend that developers and decision-makers engage critically with our findings to reduce systemic harms and avoid reinforcing dominant cultural norms at scale.

4.7 Ethical Considerations Statement

All procedures in this study were approved by the appropriate institutional review board. Participants in the video interview data collection phase provided informed consent and were made aware of how their data would be used. To preserve privacy and mitigate identity-related harms, all transcripts were anonymized by removing or replacing identifiable information such as names, universities, and geographic locations. We took additional steps to ensure that LLM evaluations did not directly expose or amplify harmful stereotypes, including careful prompt engineering and the use of controlled identity substitutions. No deceptive practices were employed. We also note that all analyses were conducted post hoc on de-identified text data, and no hiring decisions were influenced by this research. The study's goal is to surface risks and design recommendations that can contribute to more equitable and culturally inclusive AI systems.

Chapter 5

The AI Model Risk Catalog: What Developers and Researchers Miss About Real-World AI Harms

Abstract. We analyzed nearly 460,000 AI model cards from Hugging Face to examine how developers report risks. From these, we extracted around 3,000 unique risk mentions and built the *AI Model Risk Catalog*. We compared these with risks identified by researchers in the MIT Risk Repository and with real-world incidents from the AI Incident Database. Developers focused on technical issues like bias and safety, while researchers emphasized broader social impacts. Both groups paid little attention to fraud and manipulation, which are common harms arising from how people interact with AI. Our findings show the need for clearer, structured risk reporting that helps developers think about human-interaction and systemic risks early in the design process. The catalog is available at: `https://social-dynamics.net/ai-risks/catalog`.

5.1 Introduction

Risk and harm have been central concerns in AI safety and ethics research. Researchers have worked to identify and organize these concepts into taxonomies [330, 341]. Some focus on Large Language Models and Generative AI [296, 328, 330], while others address broader AI systems [337, 341] or Artificial General Intelligence [208]. These efforts often use different definitions of risk and harm. We follow the OECD [244], which defines risk as the chance of harm, and harm as a risk that has caused damage. In some contexts, the terms are used interchangeably. The MIT Risk Repository [290] compiles risks from academic work on AI frameworks and is the largest collection of researcher-identified AI risks to date. The AI Incident Database [207] catalogs real-world harms, which have been further classified by Velázquez et al. [317].

Existing research has not explored how developers describe the risks of specific models or how those models might fail in typical user scenarios. This perspective is critical, as developers can offer grounded insight into the behavior and limitations of the systems they

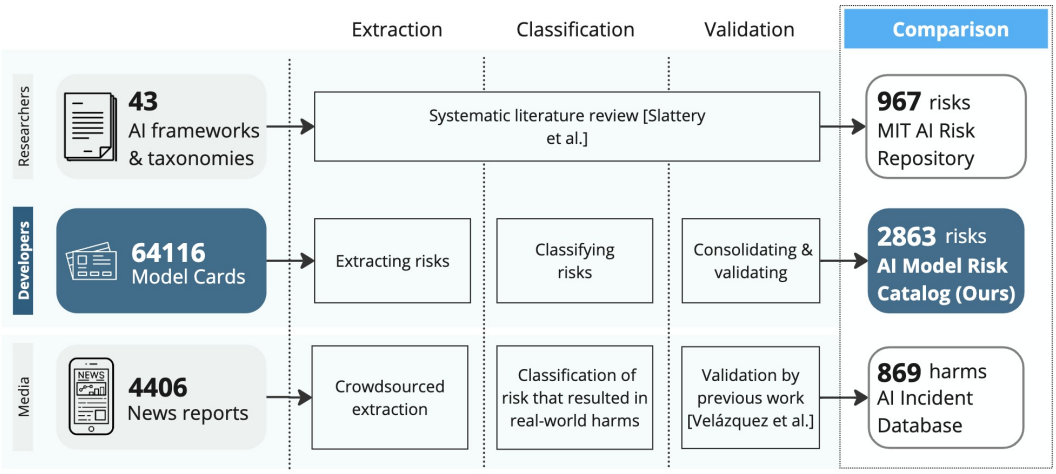


Figure 5.1: Overview of our methodology. Our methodology has three steps: extracting, classifying, and consolidating risks from model cards to build the AI Model Risk Catalog. We then compared the developer-reported risks in the catalog to those identified by researchers in the MIT Risk Repository [290] and to real-world harms from the AI Incident Database [207]. The catalog is based on 64,116 model cards with risk sections, from which we extracted 37,401 risk mentions (including duplicates). After consolidation, we identified 2,863 unique risks from 2,672 model cards, making this the largest collection of AI risks across the three sources.

build.

To address the lack of risk data linked to specific models and described by their developers, we analyzed nearly half a million model cards [213] from Hugging Face as of July 2024. Model cards are a widely adopted standard for documenting AI models, adopted by both major technology companies and individual developers. This work makes two main contributions (Figure 5.1):

1. *Analyzing risks identified by developers to create the AI Model Risk Catalog.* We collected all available model cards on Hugging Face and using established taxonomies [290, 330], grouped similar risks within each category, and selected representative examples. The result is the AI Model Risk Catalog, which includes 2,863 categorized risks tied to specific models. The catalog is publicly available at: `https://social-dynamics.net/ai-risks/catalog`.

2. *Comparing developer and researcher risks with real-world harms.* We compared the types of risks emphasized in three sources: developer-described risks from model cards, researcher-identified risks from the MIT Risk Repository, and real-world harms from the AI Incident Database. Developers tend to report technical risks such as model limitations, safety issues, and bias. These account for over half the harms recorded in real-world incidents. Researchers focus on governance, societal impacts, and threats to human agency, which together explain fewer than 15% of incidents. Notably, the largest share of incidents involves malicious use and misinformation. These risks are underrepresented by both developers and researchers, likely because they depend on unpredictable human behavior: hard for researchers to anticipate, and even harder for developers to account for [317].

We close by discussing how our catalog complements existing efforts to map AI risk and what it means for developers, researchers, journalists, policymakers, and the public.

5.2 Related Work

We start by reviewing research on AI risk taxonomies, followed by that on collating AI risks into databases.

5.2.1 AI Risk Taxonomies

The taxonomization of AI risks has grown significantly in recent years [191, 286, 296, 341]. One of the most widely used taxonomies proposed by DeepMind [330] categorizes risks into six primary groups: *representation and toxicity, misinformation, information and safety, malicious use, human autonomy and integrity,* and *socioeconomic and environmental* harms. Each category is further divided into a total of 21 subcategories, such as *toxic content* under the first category and *environmental damage* under the last. Weidinger et al. [329] also explore the context of harm origins, identifying three layers: *capability* (related to the system's technical features), *human interaction* (linked to users' experiences with the system), and *systemic impact* (including the broader context in which the system operates).

Other notable taxonomies include that by Shelby et al. [286], which identifies five key harm categories: presentation, allocative, quality of service, interpersonal, and social system harms. Yampolskiy [341] proposes a classification based on the causes of risks, organized along three axes: entity (human or AI), intentionality (intentional or unintentional), and timing (pre-deployment or post-deployment). Arguing that many existing taxonomies, including those mentioned above, are primarily designed for researchers and policymakers, making them less accessible to a broader audience, Abercrombie et al. [1] introduced the AI, Algorithmic, and Automation Harms Taxonomy, which aims to be more comprehensible to the general public. This framework adds categories such as physical and psychological harms.

5.2.2 AI Risks Data Sources

Another important area of research focuses on collating AI risks into data sources.

Risks Envisioned by Researchers. In the most comprehensive effort to date, Slattery et al. [290] analyzed 43 scholarly and industry frameworks proposing various AI risk taxonomies and compiled them into the MIT AI Risk Repository. This continuously updated repository currently lists 967 risks, categorized using two taxonomies: a causal taxonomy adapted from Yampolskiy [341] and a domain taxonomy extended from the DeepMind taxonomy [328]. Using a best-fit approach [61], Slattery et al. [290] refined the DeepMind taxonomy by adding a seventh category, *AI system safety, failures, and limitations*, and making minor adjustments to existing categories. Derczynski et al. [87] introduced "Risk Cards" providing definitions, categorizations, and examples of harms tied to specific language models (LMs). They identified risks from scholarly literature, categorized them using

the taxonomies of Weidinger et al. [328] and Shelby et al. [286], and released these findings as the LM risk cards starter set. This set, however, is a result of a more specific effort (i.e., focuses on LMs only), and features significantly fewer risks compared to the MIT AI Risk Repository.

Automatically Envisioned Risks. Tools have also been developed that often use LLMs to generate risk ideas and help AI practitioners anticipate risks during the design phase. For instance, Herdel et al. [137] created ExploreGen, an LLM-based tool that generates potential AI uses and categorizes their risk level according to the EU AI Act. Wang et al. [324] developed FarSight, an interactive tool that supports AI prototyping by providing news articles on related incidents and helping practitioners explore potential risks and affected stakeholders. Buçinca et al. [50] introduced AHA!, a tool leveraging vignettes to describe possible harms with the help of LLMs or crowdsourcing. CardGen [188] is a pipeline that uses retrieval-augmented generation (RAG) to complete missing sections of model cards, including risk sections, by drawing on information from papers and GitHub projects. Lastly, RiskRAG is another RAG tool assisting developers in envisioning risks of their models that combines ExploreGen and curates data from other model cards and the AI Incident database [259].

AI Risks Materialized as Harms. To track cases where AI systems have caused real-world harm [52], several databases of AI incident reports have been developed [268]. The most well-known is the AI Incident Database [207], which curates news reports on AI failures and categorizes these incidents [1, 309].

Velázquez et al. [317] classified incidents recorded in the incident database using the DeepMind taxonomy [330], finding that most incidents stemmed from the human-interaction layer, with harms most frequently falling under the human autonomy and integrity, or representation and toxicity categories. Bogucka et al. [39] visualized data from the incident database, including a subset focused on mobile computing [40], making the information more accessible to broader audiences. Other prominent databases include the AI, Algorithmic, and Automation Incidents and Controversies [248], the OECD Monitor [238], and "Where in the World is AI?" [6].

Additionally, AIES and FaccT communities have significantly contributed to studying instances of real-world harms [9, 11], and offered recommendations for auditing AI systems [253].

5.2.3 Research Gap

Previous work has focused on classifying AI risks into taxonomies, identifying risks envisioned by researchers, automatically suggesting risks to AI practitioners, and curating harms in the aftermath of real-world incidents. However, the valuable perspectives of AI developers have been overlooked, limiting our understanding of the unique risk profiles of specific AI models and the interactions between their capabilities and common uses. It also means we lack a clear picture of how different experts (e.g., researchers and developers) think about AI risks, and how their concerns match the harms that have already occurred.

Table 5.1: Statistics of two HuggingFace model card snapshots that we used for creating two versions of the AI Model Risk Catalog (2022 and 2024). Out of all the model cards, very few have a completed risk section, and an even smaller number have unique risk sections (i.e., sections that are not copied from other cards).

	Stats for all model cards			Stats for the model cards with unique risk sections	
Model Cards Dataset	Total cards	With risk sections	With unique risk sections	Average # downloads	Characters in risk sections
Up to 2024	461,181	64,116	2,672	118,508.41	759.69
Up to 2022	32,111	5,546	322	364,639.04	882.90

5.3 Methods

To address this research gap, we built and validated the AI Model Risk Catalog based on risks identified by developers, and compared it with risks identified by researchers in the MIT Risk Repository, as well as with harms reported in the AI Incident Database.

5.3.1 Methods for Building the AI Model Risk Catalog

We first describe the model cards dataset, and then the methods of using LLMs for extracting risks from this dataset, classifying the extracted risks into taxonomies, and consolidating and validating the classified risks into a catalog.

5.3.1.1. Downloading Model Card Snapshots from HuggingFace. In July 2024, we obtained a snapshot of the model repository from HuggingFace using the HF Hub API[39]. This included 765,973 model repositories, with 461,181 (60%) model cards. For every model card, we employed regular expressions to identify sections related to risks. Specifically, we searched for mentions of risks, limitations, bias, ethical considerations, out-of-scope uses, misuse, responsibility, and safety. This approach identified 64,116 (14%) model cards with risk-related sections. Due to the absence of standardized content requirements on HuggingFace, many model cards are incomplete, and numerous risk sections are only slightly modified replicas of one another. Among the 64,116 model cards, an overwhelming majority of risk sections (96%) were exact duplicates. We retained 2,672 model cards with unique risk-related content (selecting the card with the highest download count in cases of duplicates) as our *Model Cards up to 2024* dataset. Furthermore, we also considered a snapshot of all the model cards up to October 2022, released by previous research [184] to analyse how the risks reported in model cards have evolved over the two years. This snapshot consisted of 74,970 model repositories and 32,111 (42.8%) model cards. We extracted the risk-related sections, resulting in 5,546 (17.3%) model cards with any risk-related sections. Among these, 95% were exact duplicates with 322 model cards having unique risk-related content referred to as *Model Cards up to 2022* dataset. See Table 5.1 for details on both data snapshots.

[39] `https://huggingface.co/docs/huggingface_hub/v0.5.1/en/package_reference/hf_api`

5.3.1.2. Extracting Risk Mentions. Given the scale of our dataset, manually extracting risks from 2,672 model cards was not feasible. We therefore used a large language model (GPT-4o) to automate the extraction process and validated its accuracy against a manually annotated sample. Large language models have shown strong performance in zero-shot and few-shot annotation tasks [298, 354]. We prompted GPT-4o with definitions and examples, asking it to identify whether a section discussed risks and to extract distinct mentions in a verb–object format. To reduce variability and hallucinations, we set the temperature to zero [242]. We refined the prompt iteratively until its outputs matched human annotations on a separate 10% test set, achieving 90% agreement on a sample of 50 cards. All extracted risks were reviewed by the authors for consistency and accuracy. We removed exact duplicates and further grouped near-duplicates using two methods: (1) *fuzzy string matching* to detect similar phrasing, and (2) *contextual embeddings with cosine similarity*, using the *bge-large-en-v1.5* model [216]. Pairs with a fuzz score of at least 75 and embedding similarity above 0.85 were treated as duplicates. We retained the longer entry in each pair. These thresholds were calibrated on 5% of the dataset.

5.3.1.3. Taxonomizing Risks. After extracting risks from model cards, we classified them using two taxonomies: the DeepMind taxonomy [329] and the MIT Risk Repository taxonomy [290]. We used GPT-4o to assign each risk to the most suitable category. To guide the model, we provided definitions from both taxonomies and asked it to match each risk accordingly. We refined the prompt until its outputs matched manual annotations on an unseen 10% sample. Because the model's outputs vary, we classified each risk three times and kept only the categories that appeared at least twice. To evaluate accuracy, we compared the model's results against a manually annotated sample of 50 risks and found 84% agreement.

5.3.1.4. Consolidating Risks. A manual validation of a sample of extracted risks revealed that the automatic duplicate removal approach was partially effective. To enhance the dataset's credibility, the extracted risks were manually reviewed and refined. For each sub-category of the MIT risk taxonomy, the first author assessed the risks within the category, identified those with significant similarity, and removed duplicates, prioritizing the retention of the risk with the most comprehensive information. These eliminations were then independently verified by the second author. Any disagreements were resolved through discussion. Care was taken to preserve granular details, such as risks specific to particular models or datasets, as well as risks that represented instances of other risks but included additional information. This process resulted in our AI Model Risk Catalog.

5.3.1.5. Validating Classification of Risks. To validate the quality of the risk classifications generated by the LLM, we evaluated its performance on the manually coded MIT risk repository. We applied our custom prompt to extract the LLM's classifications of risk mentions in the repository. These predictions were then compared against the repository's manual coding, which served as the ground truth. The results showed an accuracy of 83% and a macro-averaged F1 score of 81% for the seven-class classification task. Most misclassifications occurred in the predicted categories of *malicious actors and misuse* (32 risks),

misinformation (26 risks), and *AI system safety, failures, and limitations classes* (22 risks). Upon closer inspection, we found that these were not necessarily errors. As noted by Weidinger et al. [329] and Slattery et al. [290], these categories are not mutually exclusive, and many risks span multiple categories due to their interconnected nature. For example, risks "the demonstrated ability of anonymous actors to accumulate resources online (e.g., Satoshi Nakamoto as an anonymous crypto billionaire)" and "this is the risk posed by an ideal system if used for a purpose unintended by its creators" are classified by LLM under *malicious actors and misuse*, but are found under *AI system safety, failures, and limitations* in the repository. Obviously, the unintended, anonymous, and negative use mentions also justify their inclusion under the *malicious uses* category. Similarly, "AI-generated or synthesized content can lead to the spread of false information, discrimination and bias, privacy leakage" could be both a risk of *misinformation* (as classified by the LLM) and *discrimination and toxicity* (MIT groundtruth).

To ensure that the high classification accuracy translates to our data, we conducted a manual review of how all risks are classified in our catalog, paying special attention to the overlapping categories discussed above. The review revealed less than 1% misclassifications, validating the accuracy and quality of our catalog.

5.3.1.6. Thematically Analyzing Risks. We conducted a thematic analysis [45, 46] of the extracted risks using inductive coding, where one author coded the data to comprehend and highlight the characteristics and quality of risk reporting. These codes were then jointly discussed by two authors and resolved for any disagreements.

5.3.2 Methods for Comparing Risk Sources

We downloaded two prominent sources of AI risks and implemented methods for comparing our catalog to them.

5.3.2.1. Sources of Researcher-Identified Risks and Real-World Harms. We used two public sources to compare developer-reported risks with those identified by researchers and with real-world harms. First, the MIT Risk Repository[40] [290] compiles a structured repository of 967 risks drawn from 43 AI risk frameworks, categorized across two taxonomies. Second, the AI Incident Database (AIID)[41] catalogs cases where AI systems have caused harm or failed in practice. Reports are submitted by contributors, reviewed by volunteer editors, and grouped by incident across multiple media sources. As of January 2025, it included 869 incidents based on 4,406 media reports. For brevity, we refer to these sources as the repository (MIT Risk Repository) and the database (AIID) throughout.

5.3.2.2. Classifying Risks and Harms into Taxonomies Using LLM. For AI Incidents, we followed the method of Velázquez et al. [317], classifying each incident description into our two chosen taxonomies. As with model card risks, we provided definitions of the

[40] `https://airisk.mit.edu/`
[41] `https://incidentdatabase.ai`

layers and categories from both taxonomies in the prompt. Each incident description was classified three times, and categories that appeared at least twice were kept to ensure consistency. The exact prompts are in Appendix B.3. For the MIT Risk Repository, we used the manual labels already provided for the MIT taxonomy and applied the same prompting procedure to map each risk to the DeepMind taxonomy. This process gave us common coding across all three sources (model cards, MIT repository, and AIID), letting us compare how risks envisioned by developers and researchers aligned or diverged from those shown as harms in the news.

5.3.2.3. Comparing Risk Sources. To compare the prevalence of risk categories across the three sources, we calculated confidence intervals for the difference of proportions between each category using the Miettinen-Nurminen asymptotic score method [106, p. 250]. This analysis employed the `diffscoreci()` function from the `PropCI` package in R [280]. Following prior research [274], we chose this method over Z-tests, as the boundary probabilities (close to 0 or 1) in some categories render Z-tests and their confidence intervals unreliable [5, p. 164].

5.4 AI Model Risk Catalog

We built the catalog using two snapshots of Hugging Face model cards dated through 2022 and 2024. We then examined how many risks developers report, how they report them, what kinds of risks they describe, and how their reporting has changed over time.

5.4.0.1. How Many Risks Developers Report? In the snapshot of model cards up to 2024, of the 64,116 cards with risk-related sections, 54,448 (approximately 85%) lacked substantive risk information. Most of these either simply retained the default template requesting risk details (52,983), or referred to another related card for risk content. The remaining 9,668 cards contained a total of 37,401 standardized risk mentions, including duplicates.

We removed duplicates by keeping the most downloaded model cards, resulting in a final dataset of 2,672 cards with unique risk content. From these, our extraction method produced 3,588 distinct, standardized risk mentions. The consolidation of risks (including de-duplication and streamlining) reduced the preliminary set of 3,588 risks to the final set of 2,863 risks ($\sim 20\%$ reduction).

In the snapshot of model cards up to 2022, of the 4,546 cards with risk-related sections, 367 lacked risk information. The remaining 4,179 cards contained a total of 9,645 standardized risk mentions, including duplicates. Removing duplicates resulted in 322 cards with 474 standardized unique risk mentions.

5.4.0.2. How Developers Report Risks? The thematic analysis of all the extracted risks revealed five themes related to *risk reporting practices of developers*:

1. *Repetition and redundancy*: Although expressed in many different phrasings, the risk mentions tend to cover similar ground, indicating pervasive concerns among AI developers.

Those include bias and fairness, output quality and accuracy, safety and harmful content, privacy and security, operational and technical limitations, as well as some ethical and societal implications.

2. *Ambiguity*: In line with previous research [33, 74], many risk mentions are ambiguous, lack specificity, or generally address a large type of models. Some examples are "is not immune from issues that plague modern large language models", "generates confusion", "makes mistakes", and "increases risk to users if used irresponsibly".

3. *Granularity levels and specificity*: In addition to the risk mentions such as those above that are very general, others are very capability-specific. For instance, for a general risk saying that the "model underperforms on out-of-distribution data," specific versions could be saying that it does so with particular languages, dialects, input types (e.g., "jpeg artifacts"), or for certain classification labels (e.g., emotion "fear").

4. *Interdisciplinary concerns*: The risks span across multiple dimensions of AI use, from code generation and image synthesis to language understanding and ethical decision-making, highlighting that developers understand that these challenges are not isolated but rather systemic and interconnected.

5. *Warnings for deployment*: There is an overall emphasis on caution, with many entries explicitly advising against using the models in critical or unvetted applications without robust human oversight, rigorous testing, and additional safety mechanisms. Examples include "should not be used as a substitute for professional legal advice", "should not be solely relied upon for real-time critical medical decisions."

5.4.0.3. What Types of Risks Developers Report? To understand the types of model risks developers imagine, and align with prior work [290, 328], we opted to classify the risks using established taxonomies. The DeepMind taxonomy [329]—currently the most widely cited in the literature, as noted by Slattery et al. [290]—served as a natural starting point. Additionally, we found that an augmented version of this taxonomy, developed as part of the MIT Risk Repository efforts [290], introduced an extra category (*AI system safety, failures, and limitations*) applicable to many of the risks in our catalog. Moreover, this augmented taxonomy facilitates a direct comparison between our findings and those in the MIT Risk Repository, currently the largest AI risk repository. For these reasons, we chose the MIT risk taxonomy as the second one to classify our risks with.

Appendix Figure B.2 shows how the cataloged risks map onto the MIT risk taxonomy. The most common category is *discrimination and toxicity* (44%), with 24% of all risks tied to *unfair discrimination and misrepresentation*, followed by *exposure to toxic content* (8%) and *unequal performance across groups* (8%). These risks often refer to biased training data and skewed model outputs on fairness benchmarks [177]. The next most frequent category is *AI system safety, failures, and limitations* (37%), with most of these risks falling under *lack of capability or robustness* (35%). They include reports of model bugs, inefficiencies, or errors in output quality, reflecting a primarily technical view among developers. The *misinformation* category focuses on hallucinated or inaccurate information. *Privacy and security* risks relate

to code vulnerabilities, unauthorized sharing of copyrighted material, and memorization of training data. Though less common, *malicious actors and misuse* includes concerns about models enabling illegal activity or violating human rights. The *overreliance and human agency issues* category addresses AI use in sensitive domains like healthcare and the risks of anthropomorphizing models, though it accounts for less than 2% of all risks. Similarly, the *socioeconomic and environmental harms* category—also under 2%—covers the use of computational resources, legal compliance, and broader societal impacts.

These risk patterns reflect the dominance of language models on Hugging Face (62%), which are widely known for raising concerns about discrimination and misinformation [30, 239]. However, our analysis by input and output modality, shown in Appendix Table B.1, reveals more nuances and details. Non-text inputs are more often linked to *socioeconomic and environmental harms* (22% of such risks), and *privacy and security* (20%). In total, 9% of all risks involve non-text inputs, and 17% involve non-text outputs. These risks span thousands of models and reflect a known gap in safety evaluation for non-text modalities [329]. The risks identified in our catalog could serve as an inspiration or starting point while addressing this gap, and could help guide future work in this area.

We also find that different risk categories become more prominent with *multimodal* models. For models with multimodal input, compared to other risk categories, developers report more *malicious use* (14% of that category), as well as *privacy and security* risks (13%). These findings show how risk profiles shift when looking beyond text-based models.

As shown in Appendix Figure B.3, the distribution of risks across DeepMind's taxonomy reflects similar main trends described when using the MIT taxonomy. However, because the DeepMind taxonomy lacks a category for model-specific risks—*AI system safety, failures, and limitations*—some risks in our catalog could not be clearly classified and were instead all grouped under *representation and toxicity harms*, which in this case accounts for over 62%. For more clarity and easier comparison with the MIT Risk Repository, we report the rest of our findings in the main text using the MIT taxonomy.

5.4.0.4. How Developers' Risk Reporting Has Evolved? Previous research analyzing the 2022 snapshot of model cards [184] reported that developers often struggle to complete the risk sections: less than 17% of cards had any risk content reported. To understand if risk reporting has changed since then, in Figure 5.2, we compare the 2022 and 2024 model card snapshots, and our two versions of the AI Model Risk Catalog based on those snapshots. While there is a significant increase in the total number of model cards over this period, the percentage of cards with completed risk sections has *decreased* to 14%, as has the percentage with unique risk sections ($< 1\%$). We speculate this decrease happened because newer models are often derived from or fine-tuned versions of foundational models like GPT, and may end up reusing their original risk sections. Additionally, the average length of the risk sections has also declined (see Table 5.1). Our findings not only confirm those of earlier studies that highlight developers' struggles with risk communication [33, 74, 184], but they also raise a critical concern that the situation may have even worsened over time.

In terms of the types of risks developers envision, Results in Figure 5.2 indicate a relative stability overall. The notable difference is a swap in the two most frequent categories: in 2022, *AI system safety, failures and limitations* were most prevalent, followed by *discrimination*

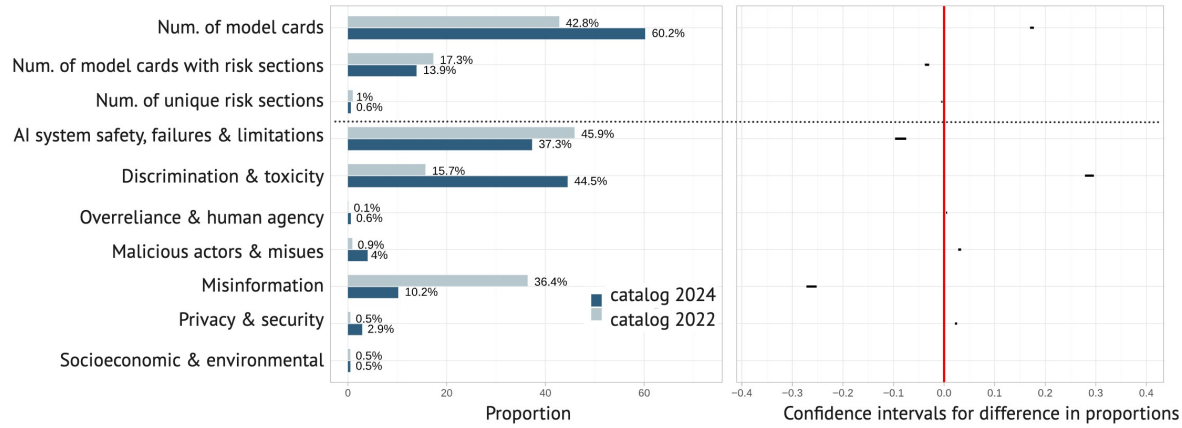


Figure 5.2: Comparing AI Model Risk Catalogs from 2022 and 2024. *Left:* The percentages of risks in the catalog that belong to each category. *Right:* 95% confidence intervals for the pairwise differences between each pair of percentages. A greater distance of the interval from the vertical red line indicates a larger difference, while a narrower interval reflects higher certainty. The top section (above the dashed line) shows statistics from the model cards snapshot, while the bottom section presents the seven risk categories. Although the number of models with model cards has grown significantly, the proportion with risk sections has dropped even further (from 17% to 14%). The *AI system safety* risks were the most prevalent in 2022, overtaken by *discrimination and toxicity* in 2024. There is a threefold decrease in the proportion of *misinformation*, and a fourfold increase in the *malicious uses* risks between the two years. The remaining categories have remained mostly stable.

and toxicity; in 2024, this order has reversed. Also, the number of *misinformation* risks has significantly reduced in 2024 (threefold decrease), whereas those about *malicious actors and misuse* have increased (fourfold increase, though still accounting for only 4% of the total risks), as well as those about *privacy and security* (sixfold increase, accounting to 3% of the total risks now). This is likely related to the increase in the multimodal models (see [346], and our Appendix Figure B.1), which are linked with more of such risks, as we discussed in the previous section. The prevalence of risks in other categories has remained relatively unchanged.

5.5 Comparison of Risks Envisioned by AI Developers and Researchers with Harms Recorded in the News

Having built the AI Model Risk Catalog, we compared risks reported by developers, those identified by researchers, and harms recorded in real-world incidents. Figure 5.3 shows the results. We highlight three key takeaways next. Note that, to capture the full range and frequency of developer-reported risks, we analyzed all 37,401 mentions, including duplicates, from 64,116 model cards with risk sections, rather than limiting our analysis to the 2,863 unique entries in the catalog. This approach aligns with the MIT Risk Repository and the AI Incident Database, where the same risks or harms may appear in multiple research papers

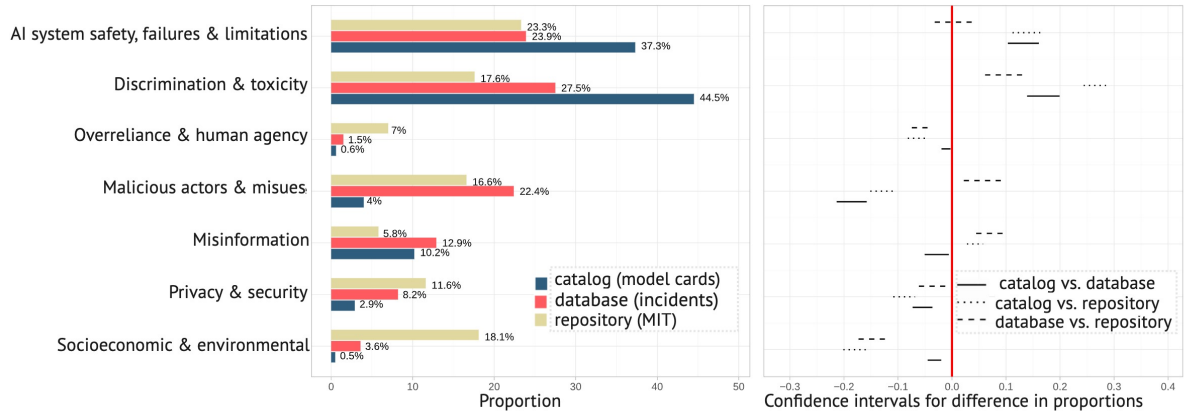


Figure 5.3: Comparison of three risk sources. We compare our catalog (developer-envisioned risks in model cards), repository (researcher-envisioned risks in the MIT Repository), and the database (harms recorded in the AI Incidents). *Left:* The percentages of risks in the catalog / database / repository that belong to each category. *Right:* 95% confidence intervals for the differences between each pair of percentages. A greater distance of the interval from the vertical red line indicates a larger difference, while a narrower interval reflects higher certainty.

or news reports.

Takeaway 1: *Developers imagine more of the risks tied to AI capabilities (37%) and bias (over 44%) than their share in real-world recorded harms (24% and 27%). Still, these risks account for over half of the harms, which supports the relevance of developers' concerns.*

When developers focus on risks tied to model capabilities (*AI system safety, failures and limitations*, 37%), it reflects their hands-on role: they build, test, and evaluate model performance. They tend to report concrete technical issues (e.g., "misclassification due to tokenization," "context window saturation," or "texts longer than 128 tokens") along with problems in how models process information, the quality of outputs, and behavior in specific settings. Unlike risks in the research repository or the incident database, the developer-reported catalog highlights fine-grained model behaviors in controlled environments. As shown in Appendix Figure B.2, most of these developer-imagined risks fall under the subcategory of *lack of capability or robustness*, which also accounts for most real-world harms. The other two subcategories—*lack of transparency or interpretability* and *AI pursuing its own goals*—each account for less than 1% of developer-reported risks and real-world harms. Still, this points to a potential gap: transparency and interpretability are key concerns in responsible AI [24, 277], especially when it comes to the adoption in high-stakes domains such as healthcare [12]. Notably, researchers give these risks slightly more attention (2.6%), as shown in the MIT Risk Repository distribution in Appendix Figure 5.3.

Bias-related risks (*discrimination and toxicity*) make up the majority (over 44%) of developer-reported risks and 27% of real-world harms. This shows that developers are not only concerned with model capability, but also with value alignment, especially around fairness and the risk of producing discriminatory or toxic content. As discussed earlier (see Figure 5.2), these concerns have more than doubled since 2022, showing that developers are paying increasing attention to discrimination from their models. These risks are often

tied to specific model types and modalities, such as speech-to-text or audio-visual models. Developers also anticipate use-specific risks: bias in models used in healthcare, hiring, or other sensitive areas is commonly mentioned. Some catalog entries are highly specific; for example, inherited bias from a certain dataset or "biases toward anime female characters." Among the three bias-related subcategories (Appendix Figure B.2), most developer-reported risks fall under *unfair discrimination and misrepresentation* (24%), followed by *exposure to toxic content* (12%) and *unequal performance across groups* (8%).

Since 24% of recorded harms relate to *AI system safety* and 27% to *discrimination and toxicity*, the concerns raised by developers are valid and timely. One example of the first type of harm involves a warehouse robot that ruptured a can of bear spray and injured workers (incident 2). An example of the second type is when Google Image search showed a Barbie doll as the first female "CEO"—after 11 rows of male CEOs (incident 18).

Takeaway 2: *Compared to recorded real-world harms, AI researchers imagine more of the risks from three categories: socioeconomic and environmental harms (18%), overreliance and human agency (7%), and privacy and security (12%). These risks appear nearly three times more often in research than in incident data, and over ten times more than in developer-envisioned risks.*

Only 4% of real-world harms fall under *socioeconomic and environmental harms*, compared to 18% of researcher-imagined risks. We see this strong focus by researchers not as misplaced but as forward-looking. As noted by Velázquez et al. [317], systemic harms often take longer to materialize and are less likely to be reported in the news—not because they are less serious, but because their effects are slower and harder to trace. Researchers give roughly equal weight (Appendix Figure B.2) to subcategories such as *power centralization and unfair distribution of benefits* (4%), *increased inequality and decline in employment quality* (3%), *economic and cultural devaluation of human effort* (2%), and *environmental harm* (3%). By contrast, developers imagine all these risks rarely, between 0.1% and 0.3%.

For *overreliance and human agency*, researchers envision 7% of their reported risks in this category, while it makes up only 2% of real-world harms, and 0.6% of developer-imagined risks. Researchers focus on both *overreliance and unsafe use* (4%) and *loss of human agency and autonomy* (3%). These risks may signal early signs of systemic issues—for example, overreliance leading to new types of mistakes at work, or loss of agency affecting worker well-being. Hence, once more, we see researchers' attention to these concerns as justified and forward-thinking.

The higher share of *privacy and security* risks in research is also notable (12% compared to 8% of real-world harms and only 3% of developer-imagined risks). While researchers envision equal proportions (6%) of *AI system security vulnerabilities and attacks* and *compromise of privacy by leaking sensitive information*, it is the latter category that results in more incidents than the former (7% *vs.* 1%).

Takeaway 3: *Both researchers and developers tend to overlook risks tied to malicious use and misinformation. Harms of this kind linked to human interaction and social engineering are significantly more common in incidents than in expert reports, indicating blind spots among the experts.*

Malicious uses of AI are widespread, accounting for over 22% of recorded harms in the incident database, yet they receive less attention from researchers (17%) and especially developers (4%). The most common subcategories are *fraud, scams, and targeted manipulation* (15%) and *disinformation, surveillance, and large-scale influence* (7%). One example of fraud and manipulation is the use of deepfake videos to scam Canadian immigrants out of thousands of dollars (incident 876). A case of political disinformation involves a deepfake that falsely showed U.S. Congressman Rob Wittman endorsing military support for Taiwan's Democratic Progressive Party ahead of the 2024 election (incident 876). These harms are often tied to specific model features [63, 124] and their data sources [190]. Hence, the low number of developer-identified risks in this area shows a clear blind spot. One example from the catalog is the risk "outputs realistic faces," linked to a text-to-video model. This risk could cause the above-mentioned harms. But this description is broad and vague—it could apply to many harms and is not helpful for mitigation. A more detailed researcher-identified risk notes that "GenAI can produce images of people that look very real, as if they could be seen on platforms like Facebook, Twitter, or Tinder. Although these individuals do not exist in reality, these synthetic identities are already being used in malicious activities." However, as with the developer risk, there are many possible misuses of GenAI that this risk could enable, which likely explains the gap between the proportions of actual harms and expert-envisioned risks in this category.

Although there is a growing body of research on *misinformation* [67, 117, 187], the actual harms seen in the world (13%) suggest that these risks are more serious and more frequent than the developers (10%), and especially researchers (6%) envision. Harms of both of these types can be linked to human interaction [317], and social engineering [281, 323] highlighting a gap in expert focus.

5.6 Discussion

By curating risks from nearly half a million model cards into an *AI Model Risk Catalog*, we help clarify how developers see AI risks. We also compare the focus of developers and researchers with real-world harms reported in the news, revealing where they overlap and differ, and where important gaps remain. Below, we outline the implications of our work for bridging the divide between these groups and other stakeholders, such as journalists, policymakers, auditors, and the wider public.

5.6.1 Theoretical Implications

Definition and format of AI risk. Our comparison of risks from two leading sources (the repository and the incident database) with those in our new catalog shows there is no standard way to report AI risks. Some risks are described in long, narrative statements that cover not just failures or incidents but also broader consequences—such as ethical dilemmas, moral challenges, or legal issues (mainly in the repository). Others are brief and sometimes vague statements about what could go wrong (often in the catalog). Real-world harms, by contrast, are tied to specific uses, domains, and subjects harmed (as in the database).

This points to a clear need for the responsible AI community to adopt a structured standard for defining and communicating risks. Drawing on insights from all three sources, we argue that a robust risk format should, at a minimum, specify the situation and context in which harm might occur. Without this, a single expert-defined risk can lead to many different harms of varying severity and likelihood (as we have discussed in *Results*). Such a standard would enable interoperability across sources and help all stakeholders—developers, policymakers, auditors, and users—understand and act on risk information. A recent step in this direction is the Risk Card format [87] for language model risks. Still, further simplification and user studies are needed to ensure risk communication meets the needs of different audiences.

Unified source of AI risks and harms. Our results show that these three resources complement each other: the repository and database provide depth of insight and examples of broader, systemic harms, while the catalog, together with risk links to specific models and datasets, can serve as a granular reference tool for the systematic auditing of AI system capabilities [312]. Together, these resources create a shared knowledge base that can support collaboration across a wide range of stakeholders.

However, while we cover two key expert groups and the media, important perspectives are still missing. These include the views of end users and affected communities, civil society organizations, and NGOs working in digital rights, consumer protection, and human rights, as well as legal experts, ethicists, and sectoral specialists from fields like healthcare, finance, and energy. Involving these groups is crucial for building a fuller, more inclusive understanding of AI risks and ensuring that risk reporting and mitigation efforts meet the needs of all those affected.

5.6.2 Practical Implications

Researchers and scholars. Our findings show that researchers have driven the discussion on the broader societal, environmental, and governance impacts of AI. However, they should place greater focus on risks linked to malicious use, misinformation, discrimination, i.e., especially those rooted in human interaction and social engineering [281]. While there is a strong body of work on misinformation [67, 187] and bias [76, 241, 332, 347], these issues remain urgent due to the frequency and severity of real-world harms, as shown by incident data. Social engineering, in particular, exploits human psychology to deceive individuals and groups, often leading to significant harm [323].

To address these challenges, researchers should develop new frameworks and practical solutions for AI risks rooted in human behavior and social dynamics. This requires closer collaboration with the human-computer interaction (HCI) community [300], and experts in social and behavioral sciences [325]. Such partnerships can help anticipate new threats of these types and design interventions grounded in real user experience.

Developers. We found that model developers, much like researchers, often overlook risks related to human interaction, social engineering, malicious use, and misinformation. However, developers also show a distinct blind spot for privacy and security risks. This lack of

attention is especially puzzling since privacy, security, and malicious use are directly tied to specific model designs, underlying data, and the amplified vulnerabilities introduced by LLMs [51, 63, 190].

Bridging these gaps will require developers to systematically consider a broader set of risks, both technical and human-centered. First, using risk taxonomies as assessment guides or "cheatsheets" can support more complete and systematic developer risk assessments. Second, we recommend drawing lessons from established cybersecurity and privacy practices and approaching AI systems as software that can be compromised. Security information sharing and the adoption of security-first and privacy-first approaches—as advocated by the NIST AI management framework [226, 227]—are key strategies [312]. Third, developers can also benefit from tools like RiskRAG [259], which guide risk thinking in relation to real-world AI uses. Future automatic tools for envisioning risks should ground suggestions in evidence from our catalog, repositories like the MIT AI Risk Repository, incident databases, and structured resources such as BenchmarkCards [293], rather than relying solely on LLMs.

Media professionals. Our findings can help media professionals reflect on how they cover AI risks and harms. Rather than focusing only on high-profile or sensational incidents, media outlets can broaden their reporting to include a wider variety of risks and affected groups. By highlighting both common and overlooked harms, media professionals can play a key role in informing researchers and the public, shaping balanced discourse (e.g., by following UNESCO's guide to best practices in AI reporting [153]), and holding developers and policymakers accountable.

Policy makers. Policymakers can benefit from these risk sources by drawing on risks identified by each of developers, researchers, and in real-world incidents when shaping legal frameworks [125]. Consolidating diverse risks in this way helps pinpoint areas needing urgent attention and can guide resource allocation. These structured resources support the development of targeted regulations, compliance tools [37, 125], and effective mitigation strategies [312].

Unlike less detailed sources, our catalog links risks to specific models, making it more useful for compliance checks and integration into compliance platforms. Our catalog also tracks how often each risk is reported, highlighting the most common concerns across different model types and modalities. Notably, as multimodal models become more widespread, they introduce new safety challenges [129, 329], and our results show these models are rapidly increasing in number among the open-source ones on the Huggingface platform—an area that requires special attention from policymakers. Even though many HuggingFace models lack risk documentation, sharing model-specific risks in a public repository like ours can aid model selection, mandatory impact assessments [37, 38], and inform product choices for end users [39]. Lastly, by adopting and encouraging open risk reporting standards, policymakers can further promote transparency and public trust in AI.

Public. Human-interaction and social engineering risks deserve special attention, and preventing these harms requires greater public awareness. The risk sources, such as our catalog, can help the public understand how AI might affect them [39], support informed choices

about adopting technology, and encourage accountability among developers, organizations, and policymakers. Public engagement in risk reporting, e.g., through participatory auditing and discussion, can strengthen accountability and improve AI integration into society.

5.6.3 Limitations

Limitations and biases of the risk sources. We treated the MIT Risk Repository as a source of risks identified by researchers, the AI Model Risk Catalog as reflecting risks described by developers, and the AI Incident Database as evidence of real-world harms. These categories are not clear-cut: many developers are also researchers, and their roles often overlap. The repository includes only a subset of academic work; most model cards do not mention risks, limiting developers' perspectives; and the incident database may reflect media coverage biases [48, 285]. While Hugging Face is one of the most comprehensive public sources of model documentation, our catalog includes only its model cards. It excludes risks described in other platforms, such as GitHub repositories or internal company records [107]. Finally, the catalog reflects a snapshot from July 2024. It does not capture model cards added or revised since then. Our findings should be read in light of ongoing updates to Hugging Face.

LLM shortcomings. Generated risks may involve inaccuracies from LLM hallucinations [214]. To mitigate this, we extracted and minimally standardized risk sentences to limit misrepresentation. Potential biases in LLM-based classifications [194] were addressed by repeating classifications thrice and applying majority agreement, with manual validation ensuring reliability.

5.7 Adverse Impact Statement

The proposed risk catalog has the potential for dual-use. Any documentation reporting failure modes and risks could be leveraged by malicious actors to scale harmful or dangerous applications of AI systems. However, the risks included in the catalog are derived from publicly available sources, i.e., model cards, and do not introduce new or undisclosed vulnerabilities. By consolidating and categorizing these risks, our aim is to support responsible AI development and risk mitigation, while minimizing the potential for misuse through careful curation and transparency.

5.8 Ethical Considerations Statement

This work adheres to ethical research principles and community norms, ensuring compliance with applicable laws and professional ethical codes. The data used in our study were sourced entirely from publicly available model cards on HuggingFace and established repositories like the MIT Risk Repository and AI Incident Database. HuggingFace allows the use, display, publication, reproduction, distribution, and creation of derivative works according to their

terms of service. We did not collect or utilize sensitive user data, nor did we interact with users or deploy systems during the research.

To mitigate ethical concerns, we took several precautions. First, we acknowledge the potential dual-use risks of documenting AI failure modes and risks, which could be exploited for harmful purposes. However, all risks in our catalog were derived from publicly accessible data, and no new or undisclosed vulnerabilities were introduced. Second, we addressed biases and inaccuracies inherent to LLM outputs through repeated classification runs, majority voting, and manual validation. Finally, we made efforts to ensure the curated catalog contributes to responsible AI development by consolidating risks in a way that supports transparency and accountability.

Chapter 6

RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting

Abstract. Risk reporting is essential for documenting AI models, yet only 14% of model cards mention risks, out of which 96% copying content from a small set of cards, leading to a lack of actionable insights. Existing proposals for improving model cards do not resolve these issues. To address this, we introduce RiskRAG, a Retrieval Augmented Generation risk reporting solution guided by five design requirements we identified from literature, and co-design with 16 developers: identifying diverse model-specific risks, clearly presenting and prioritizing them, contextualizing for real-world uses, and offering actionable mitigation strategies. Drawing from 450K model cards and 600 real-world incidents, RiskRAG pre-populates contextualized risk reports. A preliminary study with 50 developers showed that they preferred RiskRAG over standard model cards, as it better met all the design requirements. A final study with 38 developers, 40 designers, and 37 media professionals showed that RiskRAG improved their way of selecting the AI model for a specific application, encouraging a more careful and deliberative decision-making. The RiskRAG project page is accessible at: https://social-dynamics.net/ai-risks/card.

6.1 Introduction

As artificial intelligence (AI) becomes increasingly pervasive, identifying and reporting potential risks of an AI model is essential for its responsible and trustworthy development and use [90]. Ensuring AI safety is shared among stakeholders in the AI lifecycle [232]. By systematically reporting risks, AI developers can tackle ethical issues like fairness, accountability, and bias [149, 327], aligning AI models with societal values and reducing harm. Sharing risks and mitigation strategies enables application developers and organizations to enhance the safety and reliability of AI technologies. Lastly, transparent risk reporting helps consumers and technology users understand the potential risks of AI systems, enabling informed decisions

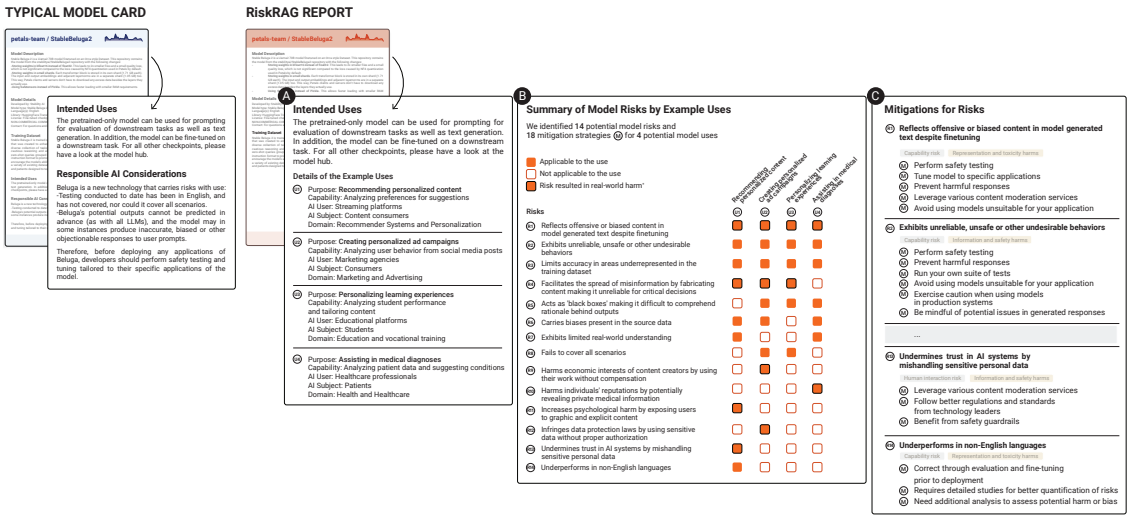


Figure 6.1: Comparison of a typical model card (left), which omits risk discussions (86% do so), with a report generated using RiskRAG (right). The RiskRAG report includes exemplary model uses (A), a summary of risks by use case (B), and corresponding mitigations (C). The report title *Summary of Model Risks by Example Uses* serves as a shorthand for *Summary of Model-specific Risks Contextualized to Example Uses*.

[113], while also addressing broader societal considerations [329], and supporting compliance with regulatory frameworks such as the EU AI Act [73] and the U.S. AI Bill of Rights [145].

Model cards have become a de facto standard for reporting on AI models, widely adopted by both major tech companies and individual AI developers alike. For instance, HuggingFace[42], one of the most popular model repository platforms, hosts 750K AI models, with 450,000 of them featuring model cards. Initially proposed by Mitchell et al. [213], model cards were introduced to standardize ethical reporting, clarify intended uses, and document the risks and limitations of AI models. Model cards encompass sections with technical information like model description, model usage, training and evaluation details, version and license, as well as the sections on intended uses, ethical considerations, risks and limitations, out-of-scope uses, and misuse.

Despite the intents of model cards, research reveals that only 17% of model cards briefly address issues related to bias or ethics [33, 184]. Our analysis of a more recent snapshot from July 2024 of 450K model cards finds this to be 14% (i.e., 64K model cards with risks reported). Moreover, a staggering 96% of these cards have their risk content copied from (i.e., identical to) an initial set of 2672 cards. Furthermore, most model documentation is insufficient for reasoning about the impact of model adoption, as the risk sections in these model cards are often criticized for being overly vague and generic, which restricts their practical application in decision-making processes [33, 74]. It has also been shown that anticipating the risks of an AI system or a model is a hard task even for practitioners and researchers with knowledge of AI [44, 98]. The increasing frequency of real-world AI incidents and harms [207, 317] is likely partly due to the lack of transparency regarding risks associated with models deployed [33, 62]. Risks can be reported as "model-specific", i.e., risks arising from the model's unique capabilities or limitations (e.g., perpetuating harmful biases from

[42] `https://huggingface.co/`

training data), or "use-specific", i.e., contextualized risks tied to specific applications (e.g., biases affecting attendees when transcribing a virtual meeting).

Two key parallel areas of research attempt to address these challenges (Table 6.7). The first area focuses on enhancing *AI risk documentation* and developing supportive toolkits. Existing studies have investigated approaches to improve the usability and effectiveness of model cards, such as introducing interactivity [74], adopting nudging formats like DocML [33], and treating AI documentation as a formal project deliverable [62] to encourage better documentation practices. These efforts have contributed to improved reporting of model-specific risks, particularly technical and capability-related risks. However, they often fall short of contextualizing these risks for specific uses [263]. This gap is becoming increasingly significant as legal frameworks, such as the EU AI Act, prioritize evaluating AI systems within their usage contexts [73]. To complement model documentation, researchers have introduced Risk Cards [86], a novel documentation format. While Risk Cards emphasize the importance of contextualizing risks for specific uses, they are designed to document individual risks separately. As a result, representing the full spectrum of risks for a single AI model would require multiple Risk Cards. Additionally, Risk Cards explicitly discourage documenting specific models, making them unsuitable for comprehensive model-level risk assessment.

The second area of research focuses on developing *tools for populating proposed documentation*. These tools are designed to help practitioners envision potential uses [137] as well as identify risks and harms [49, 324] associated with AI systems. However, existing tools do not consider the specific model underlying the AI system and therefore do not generate model-specific risks, i.e., those unique vulnerabilities or limitations tied to a particular model's design, development, or deployment. For instance, while these tools might identify risks associated with text generation language models in general, they fail to distinguish risks unique to different models that could arise, for example, from the specific data they were trained on (e.g., not safe for work images). For model cards in particular, a retrieval-augmented generation (RAG) [183] solution called CardGen [189] has been proposed to assist in filling them in. However, CardGen adheres to existing textual formats when populating risk sections, reinforcing the same shortcomings that have been criticized [33, 184] rather than addressing them [33, 74].

To address the challenges of AI model risk reporting, we built upon and contributed to both areas of research. In doing so, we made three key contributions:

1. *Enhanced AI Model Risk Report (§6.3).* Model reporting, including risk reporting, is important for a variety of stakeholders in the three AI phases of development, deployment, and use [213], as it facilitates effective communication among groups with diverse roles [127]. The AI developers who typically create model cards are our key target audience. We identified five key design requirements for AI model risk reporting through a literature review and an iterative co-design study with 16 experienced AI developers. These requirements are: identifying diverse model-specific risks, clearly presenting risks, prioritizing them, contextualizing for real-world uses, and offering actionable mitigation strategies.

2. *Automatically Pre-Populating the Report with RiskRAG (§6.4).* We developed a so-

lution to support developers in producing actionable and understandable model risk reports that align with the previously identified five design requirements. Our approach is data-driven, leveraging existing databases of human-written risks and limitations in AI models (i.e., those reported by developers in model cards, and those reported by media professionals in AI incident reports). Specifically, from an initial 450K model cards from HuggingFace, we compiled a dataset of 2672 cards containing unique risk sections and incorporated 649 real-world AI incidents from the AI Incident Database [207] to capture a diverse spectrum of risks. RiskRAG utilizes a RAG framework to retrieve relevant risks and mitigation strategies from these sources, presenting them in a clear and structured format (Figure 6.1 shows an example).

3. *Empirically Validating the RiskRAG Report (§6.5).* We first conducted a baseline evaluation (§6.5.1) to validate the quality and relevance of RiskRAG content against existing high-quality model cards written by developers. Next, we ran a preliminary user study (§6.5.2) with 50 AI developers tasked with advocating for the adoption of an AI model in a high-risk hiring application. Participants preferred using the RiskRAG report for this task compared to a baseline model card (74%), and rated it higher in meeting all the design requirements. Although AI developers typically create model cards, these reports are consumed by a broader audience, including non-technical users. Hence, in the final user study (§6.5.3), we also involved non-technical users to assess RiskRAG's broader relevance and its lower bound performance for general applicability. This study involved 38 developers, 40 designers, and 37 media professionals, who were tasked with selecting between two AI models for media industry tasks. Across all groups, RiskRAG improved the argumentation in the model selection explanation and encouraged more cautious decision-making. Participants consistently preferred the RiskRAG report, citing its clarity and support in decision-making.

Our solution improves risk reporting by providing: *(1)* an enhanced format, and *(2)* reducing the effort required for developers to document high-quality risks in this format. Importantly, we see RiskRAG *not* as a definitive solution but as *a tool to assist AI developers in creating effective risk reports.*

6.2 Related Work

Two parallel areas of related research on AI risk reporting (Table 6.7) focus on: *(1)* developing formats for documenting AI risks (§6.2.1), and *(2)* creating automated tools to assist in filling in such documentation (§6.2.2).

6.2.1 AI Risk Documentation

Various forms of documentation have been proposed to support responsible AI (RAI) practices: from model documentation (e.g., model cards [213]), dataset documentation (e.g., datasheets for datasets [119], data statements for NLP [29]), documentation for the purpose of using AI (i.e., ethics sheets for AI tasks [215]), to recent documentation for AI risks (i.e., Risk Cards [86]).

Mitchell et al. [213] introduced model cards for transparent model reporting, covering ethical considerations like sensitive data, potential risks, unintended uses, and mitigation strategies. Over time, model cards have become a standard, endorsed by regulations [73], governance frameworks [287], and major tech companies [127].

Challenges. However, model documentation format is still evolving, with some sections more frequently completed than others. An analysis of 32K model cards from HuggingFace revealed that only 17% of all cards and 39% of the top 100 most downloaded included sections on risks and limitations [184]. Another study reported similar findings when qualitatively analysing model cards also from GitHub and organizational websites [33]. Similarly, an analysis of dataset datasheets [119] found that while dataset descriptions are usually complete, considerations for appropriate data usage receive minimal attention [343]. Further analysis by Liang et al. [184] showed that risk sections of model cards typically address data and model limitations, focusing primarily on technical aspects. As a result, developers often find current risk sections ambiguous and lacking specificity [74], leading to a gap between what users need and what is provided [33].

Solutions. To tackle these issues, Crisan et al. [74] explored design choices for an interactive model cards version, while Bhat et al. [33] introduced DocML, a tool for improving documentation practices through nudging and traceability. Similarly, to incentivize risk reporting, Chang and Custis [62], suggested making model documentation a mandatory AI project deliverable. Additionally, Kennedy-Mayo and Gord [161] proposed restructuring the ethical considerations section to clearly outline regulatory, reputational, and operational risks. Beyond enhancing model cards, Derczynski et al. [86] introduced Risk Cards, a new type of RAI documentation specifically designed to address risks. Risk Cards are intended to complement other documentation by enabling cataloging of individual risks.

Research gap. To sum up, while previous research has explored improving AI model documentation, it has primarily focused on general practices rather than specifically addressing risk reporting. The efforts that focused on risk reporting, such as [161], still fall short of contextualizing risks for specific uses. We argue that only when contextualizing them to uses do the other types of risks, such as human-interaction or systemic [317], begin to emerge. Moreover, legal frameworks such as the EU AI Act prioritize evaluating AI systems within their usage contexts [73]. Risk Cards do contextualize risks for specific uses; however, they can only serve as a complementary form of documentation rather than a substitute for model cards, as they focus on individual risks rather than models, and explicitly discourage documenting specific models in relation to those risks. To address these limitations, we derived key design requirements for an effective solution through a literature-informed co-design process.

6.2.2 Tools for Populating AI Risk Documentation

The need for reporting AI risks both at the level of models and specific uses is partly driven by standards like the NIST AI Risk Management Framework [223] and regulations like the EU AI Act [73], which mandate risk documentation based on the particular use and context [126, 148]. Consequently, various AI impact assessment reports [3, 38, 295] and cards [125] have been proposed to help AI developers prepare the required documentation, particularly

for high-risk systems.

Challenges. Filling in this documentation demands envisioning the AI system's uses and risks. Besides, AI risk assessment challenges [33, 62], AI developers often struggle to envision specific uses and identify associated risks [137, 324].

Solutions. To address this, several semi-automatic tools have been proposed. Herdel et al. [137] introduced a large language model (LLM)-based tool, ExploreGen, to support developers in envisioning potential uses and assessing the regulatory risk associated with each. Buçinca et al. [49] proposed AHA!, a tool combining LLMs and crowdsourcing that assists in anticipating potential harms and unintended consequences before developing or deploying an AI system. Wang et al. [324] introduced FarSight, another LLM-based tool designed specifically to support prompt developers working with LLMs. Bogucka et al. [39] compiled risks of various AI uses that have led to real-world harms, presenting them in a visualization appealing to the broader public. All of these tools leveraged LLMs to identify potential uses or risks for a given AI system. Lastly, Liu et al. [189] introduced CardGen, a RAG pipeline that helps fill missing sections in model cards, including the risk ones, using information sourced from respective papers and GitHub projects.

Research gap. ExploreGen [137] produces only model uses, while AHA! [49], and FarSight [324] produce use-specific but not model-specific risks. For instance, they do not differentiate risks between two image generation models, e.g., one trained on NSFW (not safe for work) images, and another on safe images. The former model warrants highlighting risks related to generating abusive, violent, or pornographic content if misused, whereas the latter may not pose such risks. As we will demonstrate, RiskRAG makes this distinction (Appendix C.5). Moreover, most existing tools rely solely on LLMs, which struggle with domain-specific or knowledge-intensive tasks due to hallucinations, and a lack of grounding in the specific domain knowledge [158, 352]. In contrast, RAG [183] combines retrieval with AI-generated responses, reducing hallucinations and enhancing task-specific accuracy without additional training [116, 240].

While CardGen [189] employs RAG, and is designed to fill missing sections of model cards on HuggingFace, including risk-related sections, it does so by replicating the existing format, and with it, its limitations identified in prior studies [33, 184]. Additionally, many models on HuggingFace lack associated research papers or repositories, limiting CardGen's effectiveness in generating risk-related content for these models. Unlike CardGen, RiskRAG populates a finer-grained risk report for all model cards, even those lacking associated papers or external repositories meeting our five identified design requirements.

6.3 Deriving Design Requirements From Literature and an Iterative Co-design Process

We derived the design requirements for effective model risk reporting by reviewing the literature to identify initial requirements (§6.3.1). We then expanded on these through a co-design study with AI developers (§6.3.2).

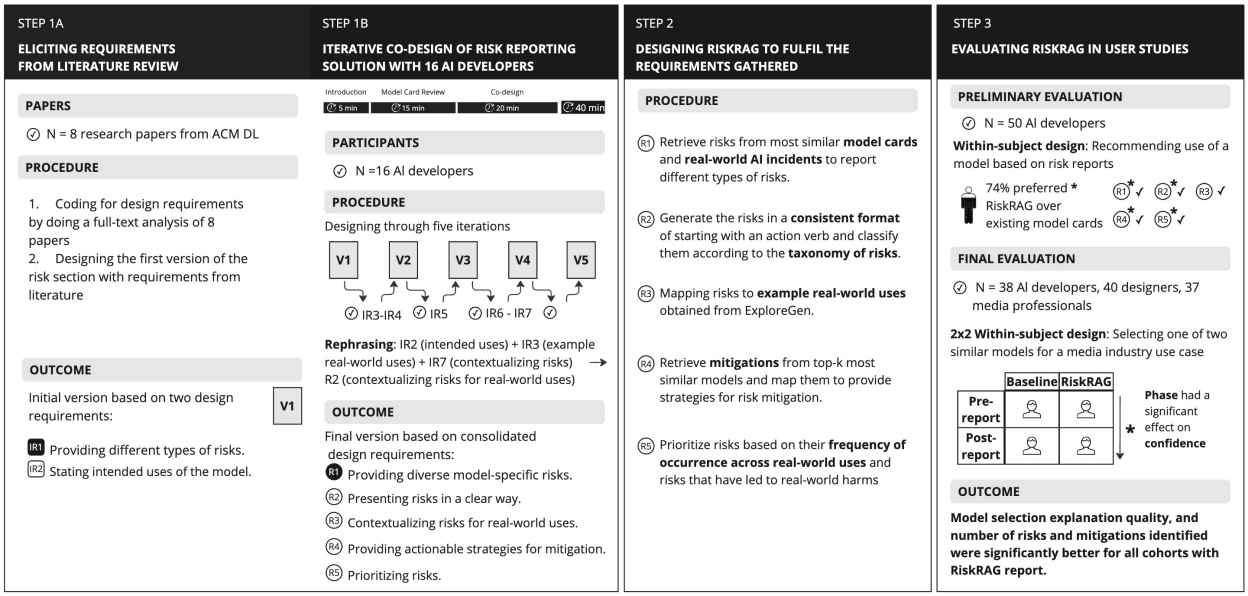


Figure 6.2: Our approach consists of three steps. *Step 1:* We identified design requirements through literature and a co-design study with AI developers (§6.3). Each co-design session included an introduction (∼5 min), a model card review (∼15 min), and a co-design task (∼20 min). After five iterations, we finalized the key design requirements. *Step 2:* We developed RiskRAG, using retrieval-augmented generation to generate risk reports aligned with these design requirements, leveraging data from model cards, and incident reports (§6.4). *Step 3:* We evaluated RiskRAG reports in two user studies (§6.5). In the preliminary study, 50 AI developers compared a RiskRAG report to a baseline model card when assessing an AI model for a high-risk hiring scenario. In the final study, 38 AI developers, 40 UX designers, and 37 media professionals compared RiskRAG reports to baseline model cards when selecting between two similar AI models for media industry tasks.

6.3.1 Design Requirements From Literature

To gather requirements for reporting risks of the AI models, we began with a recent literature review (Figure 6.2, Step 1A). Our aim was to create a foundational scaffold for a model card risk section that could be enhanced in subsequent co-design iterations. We sought to uncover major flaws in risk reporting by analyzing a selection of well-scoped, high-quality papers rather than a large set of papers identified by an exhaustive review. Prior literature has demonstrated that this approach effectively generates initial design considerations for artifact creation [38, 85].

We conducted a keyword search within the ACM Digital Library (DL) as shown in Figure 6.3, choosing this resource due to its inclusion of SIGCHI publications and proceedings from AIES and FAccT conferences, where we anticipated finding relevant papers. Our search targeted articles published from 2019 onward, marking the publication year of the first model reporting proposal, model cards [213]. The initial search yielded 326 articles. After removing extended abstracts, magazine articles, and other short-form papers or reports, 263 articles remained.[43] To ensure the relevance and quality of these papers, we applied the following

[43]Supplementary Materials (Supp. Mat.) are available on OSF at https://osf.io/chjgp/. We provide the entire list of papers in Supp. Mat. 1.1.

Search Query

Keywords: ('model cards' OR 'model documentation' OR 'model reporting')
AND
('risk*' OR 'ethic*' OR 'harm*' OR 'bias*' OR 'abuse*' OR 'misuse*')
AND
('AI' OR 'artificial intelligence' OR 'ML', OR 'machine learning')

Figure 6.3: Search query used for the literature review within the ACM DL repository.

inclusion criteria: *(1)* relevance to AI model risk reporting, *(2)* focus on the documentation practices of AI models, and *(3)* presentation of tools for documenting AI models.

Based on these criteria, we screened papers by title and abstract. This eliminated the majority of the papers focused on evaluating the limitations of the model or domain-specific papers like healthcare and education, narrowing our selection to six papers. Given the rapid development of research on ethical, responsible, and trustworthy AI, we also conducted a similar search in Arxiv to capture the latest studies, which added two unpublished papers to our list, both fairly referenced by other research attesting to their quality. The final selection of eight papers [33, 62, 74, 103, 161, 184, 213, 229] provided insights into the diverse aspects of AI risk reporting in model documentation, allowing us to synthesize initial design requirements for the risk report.[44] We conducted a full-text qualitative analysis of these eight papers. A thematic analysis [45, 46] was employed, focusing on sections relevant to risk reporting to identify key design requirements for reporting AI model risks. The first two authors distributed the papers between themselves and reviewed them to gain familiarity with their content. Using a bottom-up approach, we then coded the different sections from each paper, refining them as the coding progressed. This procedure resulted in 20 codes, which were organized into a thematic map and grouped into six sub-themes. The codebook that we generated is provided in Supp. Mat. 1.2. Finally, these led to two initial design requirements (IR) (Figure 6.2, Step 1A):

IR1. Providing different types of model-specific risks. Risk reporting of models should include different types of potential risks associated with model usage, including data and model limitations.

IR2. Stating intended uses of the model. Risk reporting of models should include intended uses of the model along with out-of-scope uses and misuse, as all these are related to risk reporting.

We used these requirements to generate the initial risk report, which was designed to guide AI developers through a step-by-step process to evaluate their models' intended uses, assess model risks, and also identify potential gaps.

[44]Refer to §6.2 for a review of these papers. Appendix C.1 lists these papers and their relevance to risk reporting in model cards.

Table 6.1: Demographics of AI developers (P1-P16) who participated in our co-design sessions. Our participants have diverse expertise with practical, hands-on experience in deploying AI models across real-world applications. M-Male, F-Female; Experience in AI is mentioned in number of years.

Version	ID	M/F	Age	Education	Expertise	Expr. in AI	Role	Hands-on Experience
V1	P1	M	25	MSc	NLP	4	Data scientist	Chatbots for e-commerce applications
	P2	M	24	MSc	NLP, CV	5	Researcher	Facial expression generation for interactive user applications
	P3	F	22	PhD	NLP, CV	5	Researcher	Models for complex network analysis
V2	P4	M	22	BSc	NLP, CV	5	Software engineer	Chatbots for hotel and finance applications
	P5	M	28	MSc	CV	5	Researcher	Human face generation for virtual user applications
	P6	M	33	PhD	Recommender systems	5	Lecturer	Recommender systems and data science
	P7	M	32	PhD	CV, uncertainty quant	7	Researcher	Earth observation models for satellite data
V3	P8	M	30	PhD	NLP	4	Researcher	Smart reply systems to enhance user interaction
	P9	M	34	PhD	Privacy-preserving ML	10	Researcher	Augmented reality applications for preserving user privacy
	P10	M	29	PhD	Reinforcement learning	6	Researcher	Modeling physical activities like running
	P11	M	33	PhD	NLP, bayesian ML	5	Data scientist	Modeling marketing tasks and microscopic data
V4	P12	M	40	PhD	Data science, ML engineering	2	Researcher	Anomaly detection in ECG signals for health monitoring
	P13	F	31	PhD	ML security and privacy, NLP	7	Researcher	Diffusion models for AI safety
	P14	M	22	BSc	CV	1	Data scientist	Misinformation detection in social media
V5	P15	M	38	PhD	ML, generative AI	5	Data scientist	Personalization in retail for targeted customer offers
	P16	F	26	PhD	On-device ML	5	Researcher	Model deployments on small devices like microcontrollers

6.3.2 Design Requirements From Co-design

We conducted a series of one-on-one co-design sessions (Figure 6.2, Step 1B) with 16 AI developers (data scientists, researchers, and engineers), since they are the primary target

users of our solution. Our co-design sessions were organized into five iterations. After each iteration, a refined risk report was developed based on user feedback. Consistent with prior research [38, 85], our sessions integrated semi-structured interviews with co-design activities to enhance the study's flow and efficiency. The resulting sequence of risk report artifacts, generated after each of the five iterations, is presented in Appendix Figure C.1.

6.3.2.1. Goal The goal of these sessions was to understand AI developers' requirements for reporting risks in model cards. The aim was also to develop a refined risk report at the end of the iterative co-design process that meets the identified requirements.

6.3.2.2. Participants We aimed to achieve a diverse participant sample using snowball sampling, where the participants were asked to identify other potential subjects. We used the following screening criteria: *(1)* has graduate or undergraduate training in ML, statistics, or a related field, or has more than 2 years of experience with AI; *(2)* downloaded or uploaded a model from GitHub or HuggingFace during the past six months; and *(3)* age 18 or older. We recruited a total of 16 participants, including 13 men and three women, comprising an equal number of professionals from industry and academia. Participants brought diverse expertise, with extensive hands-on experience in developing and deploying AI models for a variety of real-world applications (see Table 6.1). Each iteration had three to four participants.

6.3.2.3. Setup. Before the session, we emailed participants with a demographic survey and a brief description of the session goals. They were also asked to provide us with a model card of a model they had used in the past six months. We prepared an initial version of a model card with only the intended uses and risk-related sections[45] based on initial requirements (Appendix Figure C.1). We selected `bert-base-uncased`[46] from HuggingFace. This model is among the top 10 most downloaded models on the repository and the second most downloaded model to have a risk-related section.

6.3.2.4. Procedure Each 40-minute session included three activities:
Introduction (5 minutes): Participants introduced themselves, described their AI/ML projects, shared their experience with platforms like GitHub and HuggingFace, and discussed the documentation practices of the models they used.
A review of a model card (15 minutes): Participants discussed the model card they brought, focusing on risk-related sections. They identified the sections they found most important for model selection, and evaluated the usefulness of the present risk-related content in anticipating potential model risks and challenges.
A co-design task (20 minutes): We introduced our version of the model card (based on the iteration) along with a task, to determine whether they would use the model for a specific high-risk use-case (i.e., a chatbot that answers questions about applicant resumes, and helps

[45]From the literature review, we found that risks of AI models were spread across different sections. Hence, we considered any of the following sections to be risk-related: intended uses, out-of-scope uses, risks, limitations, bias, ethical considerations, and responsibility and safety. For mitigations, we added a recommendations subsection.

[46]`https://huggingface.co/google-bert/bert-base-uncased`

in filtering them) and to explain their reasoning. Participants were encouraged to suggest improvements for each model card section, focusing on information that would help them better complete the task, identify missing or inadequately represented details, and refine the risk section's content and presentation. During earlier iterations, sessions emphasized understanding what participants needed to assess risks and justify model selection. In later iterations, more time was allocated to critiquing and co-designing the artifact itself.

Two authors facilitated each session: one led the questioning, while the other took detailed notes. Sessions were recorded, with participants' consent, using online meeting software. After each session, we analyzed the key issues and updated the model card according to the requirements uncovered. This revised model card was then tested in subsequent co-design sessions with the next set of participants. By the time we developed the five versions of the model card (refer to Appendix Figure C.1 for an overview of the iterations, which are described in Supp. Mat. 2.1), it became clear that our co-design efforts had yielded sufficient insights. No new significant issues were emerging, indicating that the design had reached saturation. Following established practices in cyclical action research [313], we decided to conclude the iterative process at this point (see Figure 6.1 for a quick overview, and the final version of the risk report is in Supp. Mat. 4.2).

6.3.2.5. Gathering initial design requirements from participants After each iteration, two authors conducted a thematic analysis [45, 46] of the session's transcripts, enriching them with session notes. We employed an inductive coding approach where we coded the data to comprehend and highlight the requirements and issues raised by the participants regarding risk reporting. These codes were then jointly discussed and resolved for any disagreements. These were then arranged into relevant themes to derive a list of design requirements to be addressed in the next iteration of the model card (codebook used for each iteration is in Supp. Mat. 2.2). These co-design sessions with AI developers surfaced five additional design requirements, which led to the design of a model card artifact where no further refinements were deemed necessary (Figure 6.2, Step 1B):

IR3. Providing example real-world uses of the model. Risk reporting should include example real-world uses of the model, as it can help users visualize how the model can be appropriately used and the potential risks associated with real-world scenarios. Note that this requirement differs from providing intended uses (*IR2*), which are more general (e.g., text generation tasks), while the participant asked for specific and concrete examples (e.g., the model can be used by journalists for generating news summaries). For example, P3 mentioned *"I would see like if there are some examples of the applications. Like specific applications where it can potentially have the risks. Then it will be helpful."*, while P1 explained *"It can give some practical world applications where it can be used. It just says a sequence classification...some real applications."*

IR4. Providing strategies for mitigation. Risk reporting should include recommendations or guidelines on how to mitigate the risks. For example, P4 expressed frustration with the original model card *"... because it's not fine-tuned for that specific task... even if they did give instructions [for use], it does not supply alternative solutions."*

IR5. Presenting risks in a structured and easy-to-understand way. Risks should be organized clearly and concisely, making them easy to comprehend and act upon. For instance, P7 commented *"I think sort of a clear structure [would be helpful] few people will read through like a whole text section about this thing, unless they really dive into this topic... So somehow clearly structuring that... I guess it would be bullet points or like some diagram that has a quick summary of these risks and biases and so on. And then more detailed information below... people have short attention spans."*

IR6. Prioritizing risks. Risks should be presented in a prioritized order, reflecting their impact and importance, with which they need to be addressed. As P8 noted, *"Potentially some way of just visually showing that, OK, these three out of the six have been highlighted as being major risks."*

IR7. Contextualizing risks for specific real-world uses. Risks should be clearly linked to particular real-world uses, making it easier to understand their relevance and impact in specific contexts. For instance, P3 expressed this requirement by saying *"like when I ask about my application, then if it can answer the possible risks, and it will be really helpful."*

6.3.2.6. Rephrasing the design requirements We consolidated the design requirements gathered from the literature and co-design process into five main requirements. We did this to make them more focused and orthogonal to each other, as, for example, some of the requirements were more specific versions of the other. This also enhanced clarity, making it easier to implement the requirements in practice. For instance, the three initial requirements *IR2* (intended uses), *IR3* (example real-world uses), and *IR7* (contextualizing risks) were all consolidated into a single new requirement *R3* that encapsulates all of them.

Final design requirements:

R1. Providing different types of model-specific risks.

R2. Presenting risks in a structured and easy-to-understand way.

R3. Contextualizing risks for specific real-world uses.

R4. Providing actionable strategies for mitigating risks.

R5. Prioritizing risks.

6.4 Designing a Risk Reporting Solution (RiskRAG) Based on the Requirements

To meet the identified design requirements for AI model risk reporting (Figure 6.2, Step 2), we developed RiskRAG (Figure 6.4), a Retrieval Augmented Generation (RAG) based solution. An automated solution can effectively assist AI developers by simplifying the complex task of envisioning and documenting AI model risks, ensuring consistent and thorough reporting

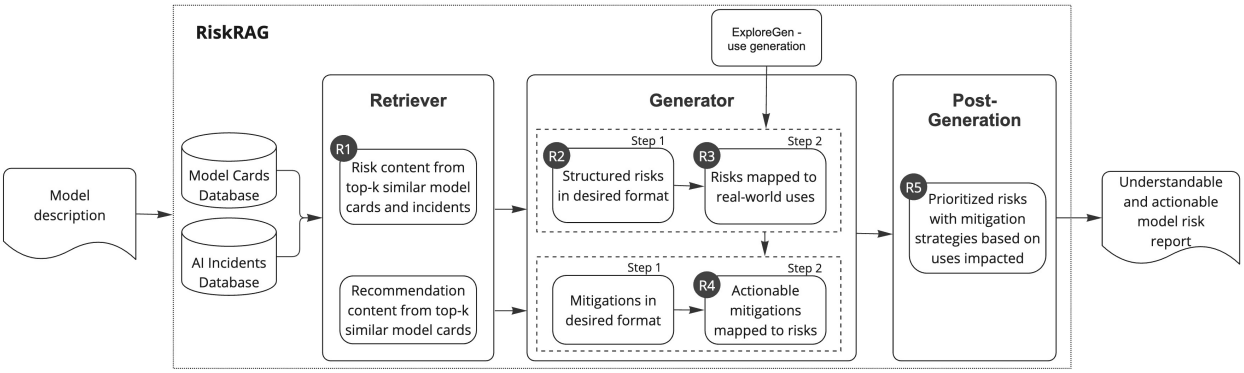


Figure 6.4: Architecture of RiskRAG. We denote with *R1-R5* different steps aiming at fulfilling the design requirements from *R1* to *R5*. The input to RiskRAG is a description of the model for which a risk report needs to be generated. The retriever first extracts risk-related content from the top-k similar model cards and AI incidents *(R1)*. The generator then adapts these risks into a standardized format and structures them using the risk taxonomy in [329] *(R2)*. The ExploreGen LLM module [137] generates examples of real-world uses to which different risks are mapped to *(R3)*. Mitigation strategies are similarly retrieved from model cards, formatted, and mapped to the corresponding risks *(R4)*. Finally, risks are prioritized based on the number of uses they were mapped to and whether they have resulted in real-world incidents *(R5)*.

across different models. RAG is well-suited for this task because it combines retrieval-based and generation-based methods, ensuring that identified risks are relevant and grounded in real-world knowledge sources. By leveraging real-world and human-written datasets with risks, RiskRAG provides a reliable solution that complements developers' expertise, making it easier to report risks that meet all design requirements gathered.

6.4.1 RiskRAG Architecture

RiskRAG uses a standard RAG architecture [183] (naive RAG as described by [116]), which combines two stages: a retrieval model and a generation model. We used pre-trained models for both, as previous work [240] has shown that this method performs well without needing extra training and generally does better than fine-tuning models with specific data.

6.4.1.1. Dataset We used two complementary datasets for retrieval: model cards (a source of both model risks and mitigation strategies) and the AI incidents database (a source of risks that resulted in real-world harms).

Model cards dataset. We downloaded a snapshot of the model repository published on HuggingFace[47] in July 2024 using the HF Hub API[48]. This consisted of 765,973 model repositories, out of which 461,181 (60%) had model cards. For each collected model card, we used regular expressions to search for risk-related sections. In particular, we searched for risks, limitations, bias, ethical considerations, out-of-scope uses, misuse, responsibility, and safety sections. This led to 64,116 (14%) model cards with risk-related sections. In the absence of standardized and strict content requirements by HuggingFace, collected model

[47] `https://huggingface.co/models`

[48] `https://huggingface.co/docs/huggingface_hub/v0.5.1/en/package_reference/hf_api`

cards were mostly incomplete, and many risk sections were only minimally modified copies of existing ones. Specifically, among the 64,116 model cards with risk-related sections, a huge majority (96%) had risk sections that were exact duplicates of another card. We further filtered out this dataset and retained 2672 model cards having unique risk-related sections (we kept the most downloaded among the cards with duplicate sections), as our final model cards dataset (see Table 6.2 for statistics).

AI incidents dataset. The AI Incident Database[49] is a publicly accessible resource that catalogues instances where AI systems have caused harm or failed in significant ways. By documenting these events, the database aims to promote transparency, improve understanding of AI risks, and guide the development of safer, more reliable AI systems. We chose to utilize this database as an additional source of risk information because it provides a crucial, real-world perspective on the various types of AI model risks that have manifested in deployments. As of March 2024, there were 649 incidents and 3412 reports, each incident derived from one or more reports. For example, one of the incident descriptions is "Meta's open-source large language model, LLaMA, is allegedly being used to create graphic and explicit chatbots [...] that participate in text-based role-playing allegedly involving violent scenarios like rape and abuse."[50] We collected the descriptions, metadata, and news reports about these 649 incidents as our AI incidents dataset.

6.4.1.2. Retriever RAG retrievers are used in tasks like question answering, where the answers must be retrieved by comparing queries to source documents using cosine similarity. We used this method to retrieve risk-related sections from similar models as well as similar descriptions from incidents by treating the model description as the query. Our source documents included both model cards and AI incident descriptions. We computed contextual embeddings for query model descriptions and source documents. We calculated similarity scores between the query and source documents to identify the top-k most similar models and incident descriptions in each dataset. Despite differences in presentation, contextual embeddings effectively capture semantic meaning across model cards and incident descriptions, as shown by prior research demonstrating their ability to handle complex linguistic structures, ambiguous word usage, and novel or domain-specific terms [15, 216, 263]. Some incidents specify model names, as the one in §6.4.1.1 involving the `Llama` model, which was matched to variants of `Llama`, as well as similar text generation models such as `falcon-7b` or `phi-2`. Other incidents lack specific model names but allow inference of the AI system's capabilities, linking them to relevant model types. For example, the incident described as "alleged AI-generated photo alteration leads to inappropriate modifications in speaker's conference picture",[51] resulted from an Image-to-Image generation model and was associated with models such as `flux-ip-adapter-v2` or `instruct-pix2pix`. We experimented with k = 5, 10, 15 to optimize for best results. From the top-k model descriptions, we took their corresponding risk-related sections. These top-k risk-related sections and retrieved incident descriptions are given as input to the generator.

An analysis of our model card dataset showed that mitigation strategies are either in a

[49] `https://incidentdatabase.ai/`

[50] `https://incidentdatabase.ai/cite/578`

[51] `https://incidentdatabase.ai/cite/820`

dedicated section like *Recommendations* or *Responsibility and Safety* or integrated within risk-related sections. Therefore, we used a combination of top-k retrieved risk-related and recommendation sections for extracting mitigation strategies.

We experimented with one sparse model: `tfidf n-gram` and three dense embedding models: `SFR-Embedding-2_R`, `Linq-Embed- Mistral` and `bge-large-en-v1.5`.[52] The dense models were selected based on their high rankings in the Massive Text Embedding Benchmark [216] (MTEB) leaderboard as of July 2024. The first one led in overall performance across 56 datasets, the second one excelled in retrieval tasks, and the third was the top performer with the fewest parameters. The `tfidf n-gram` model was included to compare the performance of traditional sparse representations against state-of-the-art dense embeddings. We used the n-gram range of 1-2.

6.4.1.3. Generator RiskRAG uses GPT-4o [101] as a generator (prompt used is in Supp. Mat. 3.1), as it is one of the leading LLMs for a variety of generation tasks [192], which also balances cost with efficiency. We devised a two-step generation for risks:

1. From the top-k retrieved risk-related sections and incident descriptions, we generated risks in the desired format of `verb + object + [explanation]`, starting with an action verb. We generated zero or more unique risks from each retrieved risk-related section, and zero or more unique mitigation strategies from risk and mitigation-related sections. Further, we classified risks along two dimensions based on the taxonomy by Weidinger et al. [329]: where they occur (capability, human interaction, or systemic), and the type of harm they represent (e.g., representation and toxicity, misinformation, malicious use). Risks generated from incident descriptions were labeled as those that resulted in real-world harm.

2. RiskRAG uses ExploreGen [137] to generate a set of realistic and diverse model uses described using a five-component format: *domain*, *purpose*, *capability*, *AI deployer*, and *AI subject* [126]. ExploreGen outputs uses across 46 varied domains. We prompted it to additionally sort these uses by their likelihood and took the top four as examples. Each generated risk was mapped to a real-world use based on its relevance to the use.

We devised a similar two-step generation for mitigation strategies:

1. From the top-k retrieved risk-related and recommendation-related sections, we generated one or more unique mitigation strategies in the same desired format as for risks.

2. Additionally, each generated mitigation strategy was mapped to one or more of the generated risks for which it was relevant.

After generating, RiskRAG prioritizes the risks (Figure 6.4, post-generation) based on how frequently they are mapped to example real-world uses, assuming that risks affecting more uses have a greater potential impact. Additionally, risks that have resulted in real-world harms in AI incident data are given a higher priority. While quantifying the impact

[52] `https://huggingface.co/Salesforce/SFR-Embedding-2_R`, `https://huggingface.co/Linq-AI-Research/Linq-Embed-Mistral`, `https://huggingface.co/BAAI/bge-large-en-v1.5`

and priority of risks remains an open research challenge [245, 262], this approach offers a simple and practical initial method for risk prioritization.

6.4.2 Meeting the Design Requirements

RiskRAG meets *R1* (different types of model-specific risks) because the retriever pulls risks from the most similar model cards, and similar models usually share comparable risks and limitations. For example, models trained on similar datasets or fine-tuned from the same parent model often exhibit similar biases, ethical issues, and fairness concerns. For instance, `bert-base-uncased`[53] and `distilbert-base-uncased`[54] (later derived from the former) exhibit similar biases related to gender and race, as noted in their model cards. In the same vein, issues related to model interpretability and robustness often arise in models with similar architectures, regardless of their specific application areas. By retrieving between $k = 5$ and $k = 15$ similar model cards, the retriever enables us to capture a substantial portion of the model-specific risks associated with the target model, which are predominantly technical and model-capability-related [263]. To capture a broader range of risks, especially human-interaction ones [317], RiskRAG also retrieves risks from real-world AI incidents linked to the model's use. For example, the ChatGPT model is associated with AI incident *642*, which describes a glitch that disrupted user interactions with nonsensical outputs.[55] Since not all incidents specify the underlying AI model, we link incidents to models performing the same task (e.g., incident *642* would also be linked to other similar text generation models). These incidents often reveal harms from human-AI interactions [317] that are underrepresented in model cards, further helping to meet *R1*. Furthermore, the generator adapts all identified risks, both those originating from similar models and those from related incidents, to the unique context of the target model. This adaptation process may involve dropping risks that are not applicable to the target model or modifying them to reflect the model's specific characteristics. For example, a risk such as "underrepresents cultures using non-English languages" might be adapted to "underrepresents cultures using non-Chinese languages, if the target model is trained on Chinese rather than English text."

RiskRAG meets *R2* (structured and easy-to-understand risks) by generating the risks in a consistent and actionable format (as described in §6.4.1). An example risk assigned to text generation models is: "undermines user trust by providing inappropriate suggestions." This well-defined format ensures that risks are articulated clearly, minimizing ambiguity. When risks are framed with action in mind, it becomes easier to implement effective mitigation strategies. RiskRAG also structures these risks using the risk taxonomy [329], further helping to meet *R2*. For instance, the example risk above is classified under the category of *information & safety harms* and in the *human-interaction* layer.

RiskRAG meets *R3* (contextualizing risks for specific uses) by mapping each risk to example uses generated by ExploreGen, contextualizing it for these specific real-world applications. For example, if the model in question is designed for text generation, then the risk "undermines user trust by providing inappropriate suggestions" is applicable for the

[53] `https://huggingface.co/google-bert/bert-base-uncased`
[54] `https://huggingface.co/distilbert/distilbert-base-uncased`
[55] `https://incidentdatabase.ai/cite/642`

Table 6.2: Statistics of our evaluation dataset for assessing RiskRAG. From the whole dataset of 2672 model cards that do not have risk content copied from each other, we took the top 10% most downloaded ones as our evaluation set.

	Evaluation set	Remaining set	Whole dataset
Number of model cards	267	2405	2672
Average number of downloads	1181494.73	497.04	118508.41
Length of risk-related sections (characters)	1233.22	707.12	759.69

use *detecting harmful content*, while its risk "violates fairness in recruitment by giving false positive results" is applicable to the use *enhancing job matching*.

RiskRAG meets *R4* (actionable mitigation strategies) by retrieving mitigations from most similar model cards and mapping them to specific risks to provide strategies for risk mitigation. An example mitigation for the risk mentioned above is: "filter the outputs of the model for irrelevant or inappropriate suggestions."

At last, RiskRAG meets *R5* (prioritizing risks) by leveraging risk frequency of occurrence across real-world uses and giving a higher importance to the risks that have led to real-world harms based on the AI incident data. For example, "violates privacy rights by disclosing sensitive personal data" is given higher priority compared to "replicates inherent biases in data" as the former was retrieved from AI incidents and resulted in real-world harm.

To sum up, retrieving the top 5 to 10 similar model cards enables the retriever to capture the broader context surrounding the target model, while the generator refines this content to address the model's unique characteristics and nuanced, context-specific risks.

6.5 Evaluating Our Risk Reporting Solution

Before evaluating whether RiskRAG produces risk reports that meet the design requirements, we first needed to determine its preliminary effectiveness in generating relevant risk content for these reports. To do so, we initially conducted a baseline evaluation (§6.5.1) of RiskRAG-generated content. Once we ascertained that our method performs well, in a preliminary user study (§6.5.2), we assessed the alignment of RiskRAG reports with the design requirements. In a final user study (§6.5.3), we evaluated whether RiskRAG can assist in selecting the most suitable AI model for a given use and support decision-making.

6.5.1 Baseline RiskRAG Evaluation

There are no standardized approaches for evaluating the RAG-generated content [305]. The challenges arise from variations in retrieved content and the common absence of ground truth for customized generation pipelines [116, 210]. In our case, there is also a lack of ground truth, as the model risk sections formatted according to all our requirements do not yet exist (e.g., risks based on uses sorted by priority). To address this challenge, we focused on evaluating the risk content format that most closely matches those available in existing model cards (i.e., individual risk/mitigation descriptions).

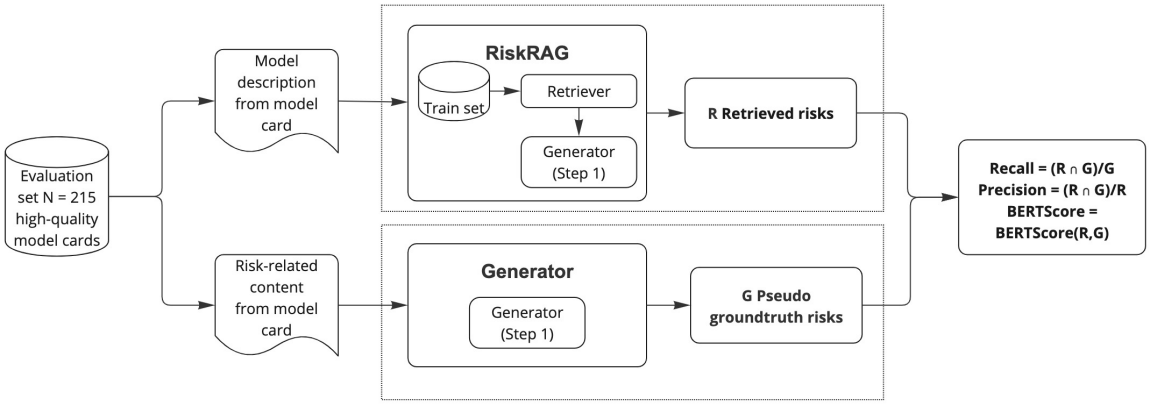


Figure 6.5: Baseline RiskRAG evaluation. This evaluation is performed on the evaluation set consisting of the top 10% most downloaded model cards (Table 6.2). We first produced risks R with RiskRAG for each of these cards using only their model descriptions. To assess the quality of RiskRAG's output against the existing risk sections of these cards, we parsed these risk sections through the generator (step 1), generating pseudo ground truth G to make them compatible with the risk content generated by RiskRAG, enabling direct comparison.

6.5.1.1. Goal The main goal of the baseline evaluation was to establish that it produces relevant risk content from which the final risk report, meeting the design requirements, can be produced. An additional goal was to determine the optimal parameters for RiskRAG, i.e., the retrieval (embedding) model to be used in the retriever, and the number of similar model cards (k) to be retrieved.

6.5.1.2. Evaluation setup One more challenge for our evaluation was that the risk sections in existing model cards are often incomplete or sparse, making traditional evaluation methods with train and test data splits difficult (e.g., we cannot include many model cards with sparse risk sections in either the training or test set.). To address this, we created an evaluation set using high-quality model cards as a pseudo ground truth. Specifically, we automatically extracted 267 model cards, i.e., the top 10% most downloaded ones (see "Evaluation set" in Table 6.2). We used popularity as a proxy for model card quality [184]. To ensure that these cards are indeed of high quality, we manually inspected a subset of 30 cards. Table 6.2 shows that, on average, these cards have longer risk reports compared to other cards, and our manual inspection confirmed that they contain non-sparse and generally carefully written risk content. Since RiskRAG generates individual risks rather than complete sections, we needed to adjust the evaluation set to match our output format. We accomplished this by processing all the risk sections from the evaluation set through step 1 of the RiskRAG generator (Figure 6.4). This resulted in a set of 215 model cards containing individual risks (even among the most popular model cards, some were found without risk content, and we dropped those), creating a pseudo ground truth (G) in our proposed format. Finally, we compared different retrieval (embedding) models and tested various values of the parameter k to identify the best performer and determine the optimal k. The complete evaluation setup is shown in Figure 6.5. To determine the hyperparameters of the system, we also compiled a separate validation set by randomly sampling model cards ($n = 25$) using

Table 6.3: Results of the baseline RiskRAG evaluation. Information retrieval (precision and recall) and the text generation (BERTScore referred to as BScore) metrics are shown. The parameter top-k represents the number of most similar model cards from which risks were retrieved.

	top-k = 5			top-k = 10			top-k = 15		
	Precision	Recall	BScore	Precision	Recall	BScore	Precision	Recall	BScore
Linq-Embed-Mistral	0.32	0.71	0.51	0.20	0.75	0.42	0.15	**0.78**	0.37
SFR-Embedding-2_R	0.32	0.69	0.51	0.20	0.75	0.42	0.14	0.76	0.36
bge-large-en-v1.5	0.27	0.60	0.48	0.18	0.66	0.40	0.13	0.68	0.35
tfidf n-gram	**0.34**	0.71	**0.53**	0.21	0.75	0.42	0.16	0.77	0.37

the same process, selecting from those that were downloaded in the top 10% to top 20% range.

6.5.1.3. Metrics We evaluated RiskRAG using two types of metrics. First, BERTScore [351] measures the overall similarity between the retrieved risks (R) and the pseudo ground truth ones (G). Second, precision and recall assess the efficiency of retrieving individual risks from the pseudo ground truth. To calculate precision and recall, we matched each retrieved risk ($\in R$) with the pseudo ground truth risks ($\in G$) using BERTScore rather than direct string matching. Direct string matching often misses contextually similar risks with different phrasing, while BERTScore, by leveraging contextual embeddings, more accurately assesses alignment between retrieved risks and pseudo ground truth. We considered a retrieved risk as correctly matched, if its BERTScore with a pseudo ground truth risk exceeded a threshold of 0.6. This threshold was determined through manual annotation on the separate validation set of randomly sampled pairs of risks (retrieved, and pseudo ground truth). For each pair, two authors evaluated whether the retrieved risk was contextually relevant to the pseudo ground truth risk. We then selected the threshold value that best aligned BERTScore matches with these manual annotations.

Recall is calculated as the ratio of correctly retrieved risks to the total number of pseudo ground truth risks $(R \cap G)/G$, while precision is the ratio of correctly retrieved risks to the total number of retrieved risks $(R \cap G)/R$. We used the original implementation of BERTScore with contextual embeddings from a pre-trained language model `DistilBERT` [279].

6.5.1.4. Results Our evaluation results are presented in Table 6.3. While there are no previous studies directly targeting our specific task, the closest approach for comparison is CardGen [189]. CardGen leverages RAG to generate model card sections based on input from related papers and GitHub repositories. The BERTScores reported by CardGen for risk-related sections, using various embedding models, range from 0.53 to 0.59, which is comparable to our results (0.53 for top-k = 5 and `tfidf n-gram`). It is important to note that our task, which requires outputting risk sections in a format different from the original

model cards, adds an additional layer of complexity to the retrieval process, making it more challenging compared to CardGen.

Table 6.3 reports precision and recall along with BERTScore. Higher recall indicates that RiskRAG successfully retrieves more risks present in the pseudo ground truth model card, while higher precision reflects fewer false positives, meaning more of the retrieved risks are indeed part of the pseudo ground truth. For top-$k = 5$, `tfidf n-gram` achieved the highest precision (0.34) and BERTScore (0.53) Both `Linq-Embed-Mistral` and `SFR-Embedding-2_R` also performed well, each with a precision of 0.32 and a BERTScore of 0.51, demonstrating their competitive accuracy and similarity. As the top-k value increases, all three models maintained strong recall, with `Linq-Embed-Mistral` slightly ahead, reaching a recall of 0.78 at top-$k = 15$. This suggests that while `tfidf n-gram` excels in precision and BERTScore, `Linq-Embed-Mistral` is slightly better at retrieving a broader set of relevant risks. Precision is comparatively low, and declines as top-k increases. However, since the risk sections in model cards in our evaluation set are likely incomplete, lower precision does not always indicate false positives. Many retrieved risks may be missing from the pseudo ground truth but remain relevant to the model as they are sourced from similar models. Additionally, Step 2 of the generator (not included in this part of the evaluation) can filter out irrelevant risks, refining the results further. Therefore, precision is less critical in our evaluation than recall or BERTScore.

To confirm that RiskRAG generates a broad and relevant set of risks despite low precision, and to resolve the tie between `tfidf n-gram` and `Linq-Embed-Mistral`, we qualitatively examined the risks for the ten most downloaded models from our evaluation set. We chose $k = 10$ for its broader risk coverage over $k = 5$, while avoiding the lengthy lists seen with $k = 15$. Indeed, an initial manual analysis on a validation set showed that $k = 10$ best balanced coverage and conciseness, given the incomplete risk sections in many model cards. Upon the manual evaluation of the ten most downloaded cards, we found that most of the risks retrieved were relevant to the model in question, and were indeed missing in the original card. For instance, the card for `google/flan-t5-large`[56], a multilingual text generation model, listed risks such as biases in training data, and harmful, inappropriate, or explicit content generation. RiskRAG expanded this by identifying additional relevant risks, including hallucination, toxicity, misinformation, unsupported languages, malicious use, and representational harms, such as racial and gender stereotypes in online data. Although these risks were not documented in the original card, they are relevant for a comprehensive risk assessment. We also observed that low precision was partly due to similarly worded risks across top-10 results and the presence of use-specific risks from similar models that did not directly pertain to the model for which risks were retrieved. These issues get mitigated in Step 2 of the generator, where irrelevant risks are dropped, and those specific ones are adapted.

We chose `Linq-Embed-Mistral` for the later user evaluation (§6.5.2) for two reasons: *(1)* Comprehensive coverage. `Linq-Embed- Mistral` provided a broader and more detailed set of risks. For DistilBERT, `tfidf n-gram` primarily highlighted biases such as "produces biased predictions despite neutral training data" and "transfers bias to all fine-tuned versions". `Linq-Embed-Mistral` included additional risks related to domain-specific performance and

[56] `https://huggingface.co/google/flan-t5-large`

language issues like "underperforms on text from different domains" and "underperforms on non-English languages". *(2)* Relevance. `Linq-Embed-Mistral` identified risks more pertinent to specific tasks. For example, it captured risks related to image classification for the OpenAI CLIP[57] model, such as "reduces performance when input images are resized" and "memorizes duplicated images in the training data", which `tfidf n-gram` missed. Overall, while `tfidf n-gram` highlighted major risks, `Linq-Embed-Mistral` offered a broader and relevant assessment, making it the better choice for detailed risk reporting.

6.5.2 Preliminary User Study

We conducted a preliminary user study (Figure 6.2, Step 3 top) to understand how AI developers perceive RiskRAG's risk reports relative to existing model card risk reports.

6.5.2.1. Goal To assess whether RiskRAG's reports meet the identified design requirements and are preferred over standard model card risk reports when deciding whether to use an AI model for a given task.

6.5.2.2. Study design We conducted a within-subject study where each participant evaluated two versions of a model risk report: a baseline version containing the risk-related sections of the original model card (control), and the model risk report generated by RiskRAG (§6.4) (treatment). Specifically, we chose two similar text generation models: `Phi-3-mini-128k-instruct`[58] (downloaded over 200K times in July 2024) and `StableBeluga2`[59] (downloaded over 130K times in the same period). These models were selected because their original model cards feature relatively rich risk sections, allowing for a fair comparison of the baseline with the RiskRAG report. Participants were then asked to consider one of two high-risk (according to the EU AI Act) uses for the models: *(1) a chatbot to evaluate and rank pre-interview assessments* and *(2) a chatbot that answers questions about applicant resumes and helps in filtering them.* We focused on high-risk scenarios because effective risk reporting is crucial in such contexts. Additionally, selecting applicants for a team is commonly experienced by developers, making these model uses relatable and relevant for our participants.

We developed a web-based survey that included a real-world task to be performed in both control and treatment conditions:

"Write a short email to your line manager asking for approval to use this model. In your email, explain the technical and ethical reasons why this model should be used, but also be candid about any potential risks and discuss how they can be mitigated."

The vast availability of AI models today makes it realistic for developers to argue for the use of any particular one and discuss trade-offs with management. In the post-COVID world, where a significant portion of work communication occurs online, as many people still choose to work remotely [264], sending an email was also considered a practical and relevant task.

[57] `https://huggingface.co/openai/clip-vit-large-patch14`
[58] `https://huggingface.co/microsoft/Phi-3-mini-128k-instruct`
[59] `https://huggingface.co/petals-team/StableBeluga2`

6.5.2.3. Metrics To assess how effectively RiskRAG met the requirements compared to the baseline risk sections from the original model cards, we measured the following (left panel of Figure 6.6):

Q1. Does the risk report provide reliable information and cover a wide range of risks? (R1)
We measured this by adapting an item for *completeness* and another for *perceived accuracy* from the AIMQ information quality assessment scale [181].

Q2. Does the risk report explain the risks in a clear and concise manner that is easy to understand? (R2)
We used items on *understandability* and *concise representation* from the AIMQ scale [181] to measure this.

Q3. Is the content of the risk report relevant to the use case presented in the task? (R3)
We adapted and measured an item on information *relevance* from the AIMQ scale [181].

Q4. Does the risk report offer clear strategies for mitigating risks? (R4)
We adapted an item on *ease of operation* from the AIMQ scale [181] to measure this.

Q5. Does the risk report effectively prioritize the risks? (R5)
We developed a custom item to measure this specific requirement that emerged during our co-design sessions.

Each question was rated on the scale from 1 (strongly disagree) to 5 (strongly agree).

The last metric we measured was the *preference* towards the baseline risk report or treatment risk report from RiskRAG. For uniformity, all metrics were measured using a 5-point Likert scale. Finally, we also had two open-ended questions: "In what ways did the report succeed in assisting you in completing the task?" and "In what ways did the report fall short of assisting you in completing the task?"

6.5.2.4. Participants. We focused on AI developers, who are responsible for assessing how AI systems perform in specific use cases, including evaluating human interaction effects and technical capabilities within their applications [329]. We recruited 50 AI developers through the online recruitment platform Prolific.[60] To ensure the suitability of participants, we applied four a priori inclusion criteria, targeting individuals who held individual contributor roles, worked in an engineering function within their organization, were employed specifically for coding tasks, and used AI multiple times per week. Additionally, we used two items from the AI literacy scale [60] to control for participants' AI literacy and their attitudes towards AI.

Participants' ages ranged from 18 to 29 (40%) and 30 to 39 (60%). All participants were male, residing in the U.S., and working as individual contributors in non-managerial technical roles, including engineering (33%), design (17%), data analysis (17%), and research and education (17%). Regarding ethnicity, 40% identified as Asian, 20% as Black, 20% as Mixed, and 20% as White. In accordance with our study requirements, 60% of participants reported using AI in their work multiple times a week, while 40% used it daily.

[60]See `https://www.prolific.com`, last accessed August 2025.



Figure 6.6: **Quantitative results from the preliminary user study:** RiskRAG report outperformed baseline model cards across all the metrics. We had seven questions mapping to our design requirements to which participants were asked to answer on a Likert scale from 1 ("strongly disagree") to 5 ("strongly agree"). RiskRAG significantly outperformed the baseline model cards, with a one-point higher rating, moving from slight agreement to clear agreement on risk coverage, and from neutral to agreement on mitigation clarity.

6.5.2.5. Procedure. The entire procedure took place in three steps. In the first step, participants answered a demographics questionnaire. In the second step, participants were provided a brief introduction to the tasks, followed by the first task in which participants had to read either the control or treatment risk report, and write the email. In the third step, they answered the questions about the chosen metrics. They then proceeded to view the other risk report and answered the same set of questions.

To counterbalance potential order effects, the sequence in which the baseline and treatment risk reports were shown to participants was randomized. To eliminate any effects from the type of AI models shown, each participant reviewed risk reports of two different models with different real-world uses, randomly assigned to ensure a thorough evaluation. Baseline and RiskRAG reports for the two models are in Supp. Mat. 4.

6.5.2.6. Analysis. After confirming non-normality with the Shapiro-Wilk test, we applied the Wilcoxon signed-rank test to assess if the RiskRAG enhancements produced statistically significant improvements over the baseline. The results of the thematic analysis of qualitative feedback from open-ended questions are detailed in Appendix C.4.

6.5.2.7. Results. Figure 6.6 presents the breakdown of the quantitative results. Across the questions, the participants rated the RiskRAG report more favorably than the baseline model card. The largest difference is evident for the questions on offering clear mitigation strategies (R4), where the RiskRAG report scored 4.08 ("agree"), and baseline card 3.08 ("neutral"), and for covering a wide range of risks (R1), where the report scored 4.46 ("clearly agree"),

Table 6.4: **Statistical significance of the results in the preliminary user study:** Wilcoxon signed-rank test results between the baseline model cards and the RiskRAG report. The test showed statistical significance across the majority of the chosen quantitative metrics, indicating that the RiskRAG report outperformed the baseline across all the identified requirements: *p<0.05; **p<0.01; ***p<0.001; *ns*:p¿0.05.

metric	z-statistic	p-value
provide reliable information (R1)	63	*
covers a wide range of risks (R1)	60	***
explain risks in a way that is easy to understand (R2)	100.5	ns
present risks in a clear and concise manner (R2)	103	*
be relevant to the use case in the task (R3)	87	ns
offer clear strategies to reduce risks (R4)	77	***
prioritize the risks effectively (R5)	149	*
preference between RiskRAG report and baseline	331.5	**

and the card 3.40 ("slightly agree"). While participants gave higher scores to the RiskRAG report for explaining risks in an easy-to-understand way (R2) (4.17 for our report versus 3.75 for the baseline) and being relevant to the use case in the task (R3) (4.27 *vs.* 4.04), the statistical tests did not show statistical significance on these questions. The differences in all the other questions, i.e., providing reliable information, presenting risks in a clear and concise manner, and prioritizing them effectively, were statistically significant (Table 6.4). The preference for the RiskRAG report over the baseline was statistically significant, with 74% favoring it.

6.5.3 Final User Study

Through the preliminary user study, we established that RiskRAG generates risk reports that fulfil all identified design requirements and are preferred over existing risk reports in model cards. Subsequently, we conducted the final user study (Figure 6.2, Step 3 bottom) to assess its effectiveness in decision-making and actionability.

6.5.3.1. Goal To assess whether RiskRAG reports improve understanding of model risks, facilitate critical evaluation, and support envisioning mitigation strategies better than standard model card risk reports.

6.5.3.2. Study design This study employed a 2×2 within-subject design. The task was to assess two models for a real-world use case, and select the one deemed more suitable. Each participant completed this task across two *conditions*:

1. **Control**: risk-related sections of the original model card; and
2. **Treatment**: RiskRAG-generated risk report.

For both conditions, the task was divided into two *phases*:

1. **Pre-report phase.** Participants chose a model after reviewing descriptions of two models and their intended use, providing an explanation for their choice.

2. **Post-report phase.** Participants reviewed a brief incident tied to the intended use and the risk reports (original model card for control, RiskRAG for treatment). They could reconsider and change their previous model choice, explaining in either case their final decision. This step assessed the impact of the risk report on their decision-making.

This design enabled a direct comparison of participants' decision-making with and without RiskRAG reports, as well as with and without baseline model card risk sections, providing insights into their different impacts on model selection.

To test RiskRAG's ability to generalize beyond text generation models that we used in the preliminary study, we selected two pairs of models, both of different types than text generation, each coupled with an appropriate real-world use and an associated incident:

1. Multimodal Models (Image or Text-to-Text): `idefics-80b-instruct`[61] and `paligemma-3b-mix-448`[62], which take both image and text inputs and produce text outputs. Use: Developing a system for a media organization to identify people and objects in photos and generate alternative text descriptions for web pages. Incident: The model mistakenly labelled a Black couple as "gorillas".[63]

2. Automatic Speech Recognition (ASR) Models (Speech-to-Text): `whisper-large-v3a`[64] and `canary-1b`[65], which process speech input and convert it to text. Use: Creating a system to transcribe spoken content into text for broadcast subtitles. Incident: The model hallucinated violent language and fabricated details, particularly during extended pauses in speech.[66]

These medium- to high-popularity models, with rich risk-related sections in their original cards, provided a strong baseline for RiskRAG. The models were selected to have comparable strengths and risks, ensuring no definitive "right" choice between them. This allowed us to focus on how participants deliberated between different pairs of models.

6.5.3.3. Metrics We measured the *explanation quality* to assess differences between the baseline condition and the treatment condition in the post-report phase. Three authors, with extensive expertise in responsible AI, human-computer interaction, computer vision, and NLP, and with a strong publication record in AI applications, risk reporting, impact assessments, and user study design, independently annotated the explanations according to a scoring rubric. Prior to performing any annotations, all evaluators participated in a calibration session to ensure a consistent understanding and application of the scoring rubric, which is detailed in the Supp. Mat. 5.1, and summarized here:

[61] `https://huggingface.co/HuggingFaceM4/idefics-80b-instruct`, last accessed August 2025.
[62] `https://huggingface.co/google/paligemma-3b-mix-448`, last accessed August 2025.
[63] `https://incidentdatabase.ai/cite/16/`, last accessed August 2025.
[64] `https://huggingface.co/openai/whisper-large-v3`, last accessed August 2025.
[65] `https://huggingface.co/nvidia/canary-1b`, last accessed August 2025.
[66] `https://incidentdatabase.ai/cite/732/`, last accessed August 2025.

1. *Number of identified risks:* Evaluators counted the number of identified risks in the explanation. These risks were either in the context of the real-world use or supported model selection, demonstrating the understanding of the report's content.

2. *Number of proposed mitigations:* Evaluators counted the number of appropriate mitigation strategies proposed to counter the identified risks.

3. *Task quality:* Evaluators rated how effectively the explanation communicated the trade-offs between risks and benefits for the selected model. This was scored on a scale from 1 to 5, where a score of 1 reflected vague reasoning, lack of argumentation, or no clear call to action, and a score of 5 reflected strong alignment with the task, robust mitigation strategies, and a well-structured argument featuring diverse trade-offs.

The inter-annotator agreement, calculated by a Fleiss' kappa, ranged from 0.82 for the number of identified risks to 0.78 for the task quality. This confirmed that the evaluators agreed with each other strongly. For each metric listed above, the final score was determined by averaging the scores given by three evaluators. The *overall explanation quality* was then calculated as the average of the final scores for all three metrics.

We measured the following *decision metrics* in pre- and post-report phases for both conditions:

1. *Decision confidence*: Participants' self-reported confidence in their model choice, measured with the question: 'How confident are you in your ability to choose the most appropriate AI model for this task?'. This was measured using a 5-point numerical scale.

2. *Decision time*: The time participants took to make their model selection.

3. *Preference*: We also measured the *preference* between reports, as used in the preliminary study (§6.5.2.3)

6.5.3.4. Participants We conducted the study in three cohorts (Table 6.5): 38 AI developers, 40 UX designers, and 37 media professionals recruited through Prolific. For the 2×2 within-subject study, we conducted a priori power analysis. A repeated measures ANOVA with medium effect size ($f = 0.40$, $\alpha = 0.05$, $1-\beta = 0.95$) indicated a minimum requirement of 15 participants, which was met and exceeded by all our cohorts. In the first cohort, we recruited developers using the same inclusion criteria as in §6.5.2.4. In the second cohort of UX designers, we recruited individuals who held individual contributor roles, worked in a design or creative function within their organization, and used AI at least once a week. The third cohort included professionals in journalism, marketing, communications, design, and creative roles, who use AI at least once a week. We selected this group because they are the primary users of the applications featured in the task. AI developers were the most knowledgeable in the task, technology, and AI across the three cohorts, and the task was most similar to their day-to-day tasks at work.

Table 6.5: Self-reported knowledge and demographic characteristics of participants in the final user study.

Control	Characteristic	AI Developers (n=38)	UX Designers (n=40)	Media Professionals (n=37)
Expertise	Task	3.97 ±0.75	3.60 ±1.08	3.65 ±1.01
	Technology in general	4.34 ±0.67	4.15 ±0.58	4.05 ±0.74
	Artificial Intelligence	4.13 ±0.74	3.88 ±0.91	3.86 ±0.75
Task similarity	Similarity with day-to-day tasks	3.34 ±0.94	2.92 ±0.94	2.92 ±1.04
Age	18-29 years	47.4%	45.0%	40.5%
	30-39 years	26.3%	22.5%	35.1%
	40-49 years	18.4 %	22.5%	10.8%
	50-59 years	5.3%	7.5%	10.8%
	60 years and above	2.6%	2.5%	2.7%
Sex	Female	26.3%	32.5%	48.7%
	Male	73.7%	67.5%	48.6%
	Prefer not to say	0	0	2.7%
Ethnicity	White	34.2%	47.5%	40.5%
	Black	34.2%	35%	35.1%
	Mixed	18.4%	12.5%	21.6%
	Asian	7.9%	5%	0
	Other	0	0	2.7%
	Not specified	5.3%	0	0

6.5.3.5. Procedure The procedure consisted of three main steps. 1. Participants provided informed consent; 2. Participants received instructions for pre-report phase, selected a model, explained their choice, and reported their decision confidence; and 3. Participants received instructions for the post-report phase, reviewed the risk report, reconsidered their selection, explained their final choice, and reported their confidence again. This process was repeated with a second set of models using the alternate risk report. To counterbalance order effects, the order of the baseline report and treatment report was randomized, and the two multimodal and two audio models were randomly assigned to either condition. Baseline and RiskRAG reports for all four models are in Supp. Mat. 5.2.

6.5.3.6. Analysis We used the Wilcoxon signed-rank test to analyse the difference in explanation quality between control and treatment. A 2×2 repeated measures ANOVA examined the effects on decision metrics, identifying main effects and interactions to assess whether RiskRAG reports significantly influenced confidence and time compared to the control. We conducted inductive thematic analysis [45, 46] to derive themes from participants' model choices and preference explanations. Following established coding procedures [212, 273], two authors coded the data for the baseline condition and report condition. The process involved familiarizing oneself with the data, iterative coding with refinement, and

Table 6.6: **Results for decision confidence from the final user study.** A 2x2 analysis based on two factors: (baseline *vs.* treatment) and phase (pre-report *vs.* post-report). Significant results are highlighted in bold. Confidence dropped significantly from before to after seeing the risk report for developers and media professionals, with a bigger drop after seeing RiskRAG compared to the baseline. *p<0.05; **p<0.01; ***p<0.001

Metric	**Factors**	**AI Developers (n=38)**	**UX Designers (n=40)**	**Media Professionals (n=37)**
Decision confidence	Condition (Main effect)	F(1,37) = 0.39	F(1,39) = 0.00	F(1,36) = 2.26
	Phase (Main effect)	**F(1,37) = 6.13***	F(1,39) = 0.09	**F(1,36) = 5.10***
	Condition x Stage (Interaction)	F(1,37) = 0.59	F(1,39) = 1.06	F(1,36) = 2.03

resolving disagreements through discussion. This analysis yielded 62 codes organized into a thematic map with 11 sub-themes and five themes for the report condition, and 32 codes grouped into six sub-themes and three themes for the baseline condition. The codebook is provided in Supp. Mat. 5.3.

6.5.3.7. Results *Explanation quality* was higher for RiskRAG reports compared to the baseline reports across all three cohorts (Figure 6.7). For developers, both the number of proposed risks and mitigations ($W = 104, p = .041$ and $W = 130, p = .021$), as well as task quality ($W = 108, p = .011$), were significantly higher. For designers, the number of identified risks ($W = 157, p = .016$) and task quality ($W = 139, p = .001$) were significantly higher. Although the differences did not reach statistical significance, the explanations of media professionals were also of higher quality.

Regarding decision metrics, Table 6.6 shows that *decision confidence* was significantly different in the post-report phase when compared to the pre-report phase for the developers' ($F(1, 38) = 6.11, p = .018$) and media professionals' ($F(1, 38) = 6.11, p = .030$) cohorts. Their confidence decreased after interacting with the reports, with a greater drop observed in the treatment condition compared to the control. Designers showed a similar trend, although it was not statistically significant. We will later examine these lower confidence scores in relation to participants' free-form comments. As we will show, participants in the treatment condition reported more cautious argumentation in their explanations, and that was linked to their decreased self-reported confidence. No significant effects were observed for *decision time*. Lastly, while we intentionally selected models of comparable quality, it was insightful to examine participants' *decision changes* with respect to to their initial model choices after they interacted with RiskRAG versus the baseline. Developers changed their mind more with RiskRAG (82% *vs.* 76%), as well designers (70% *vs.* 62%), but not media professionals (73% *vs.* 78%).

Regarding preference between the reports, the RiskRAG report was favored by 58% of developers, 63% of designers, and 70% of media professionals.

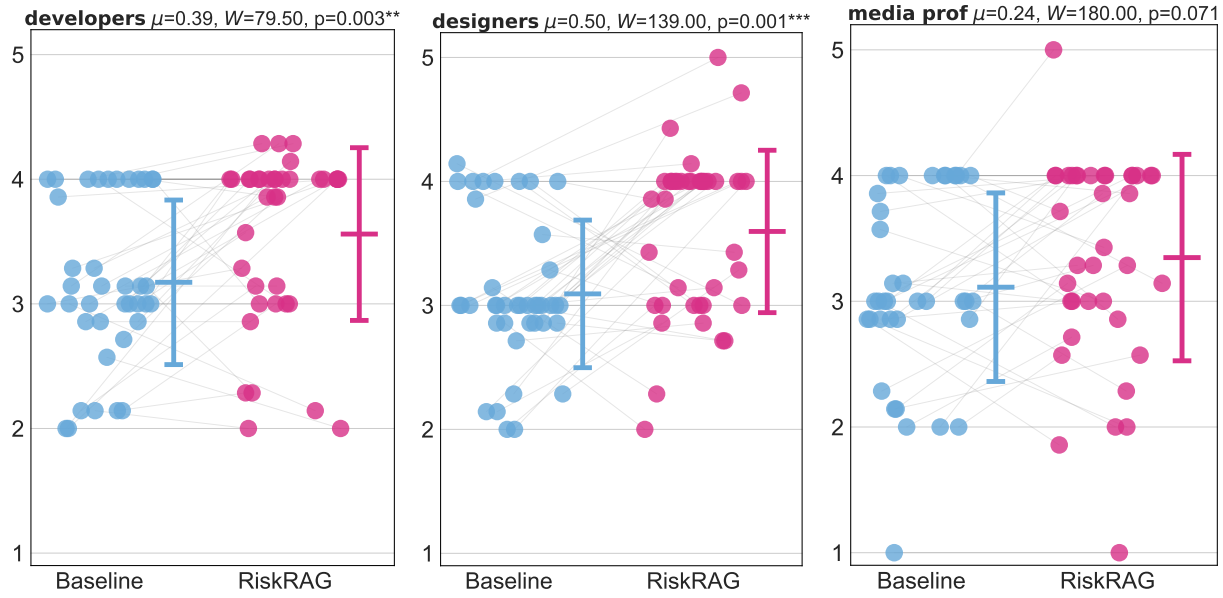


Figure 6.7: Results for explanation quality (y-axis) from the final user study in three cohorts: developers, designers, and media professionals. Mean explanation quality (μ) improved for all cohorts after using RiskRAG, as opposed to the baseline reports, with statistically significant increases as shown by Wilcoxon test results (W, p) for developers and designers. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. After using the RiskRAG report, participants identified a higher number of relevant risks and mitigation strategies and provided better quality explanations for their model selection compared to baseline.

Qualitative Results Participant preferences for the RiskRAG report revealed five key themes. Below we describe these themes along with major sub-themes.[67]

1. Supporting decision-making: In this major theme, participants appreciated the report's ability to simplify the cognitive load for analysis as it provided clear structure, prioritized risks, and revealed stakeholder impacts, enabling effective comparison across models and their decision-making. "*Risk heatmaps are effective for summarizing complex data and can be especially useful in fast-paced decision-making scenarios.*" (P0, dev).[68] "*I can understand easily what is good and what is bad, to compare and make a decision easier.*" (P11, des). A key sub-theme that emerged across all cohorts was **cautious decision-making**. Participants expressed greater awareness of potential risks, leading to more deliberate choices: *"I'll proceed cautiously with the use of MODELX, given the serious ethical and reputational risks highlighted by past incidents."* (P1, des) *"I would deliberately undertake a training to the developers and the teams that will be involved in generating the content on information sensitivity, reputation, and consistency of information."* (P4, dev) *"This model, with its larger scale and instruction-based design, has the potential to produce more accurate and nuanced description [...] However, even a large-scale model must be thoroughly tested for bias, especially in diverse contexts, to avoid making harmful associations."* (P35, med) Clear, contextualized risk reports emphasizing real-world harms led participants to approach model

[67]The full codebook with all sub-themes is in Supp. Mat. 5.3.

[68]We represent participants by their ID and cohort, dev: developers, des: designers, med: media professionals.

selection more carefully. Explicit examples of past incidents increased awareness of potential consequences, prompting them to reconsider choices initially based solely on technical details. This shift in awareness contributed to a notable reduction in decision confidence in the RiskRAG group during the post-report phase. Although both groups showed decreased confidence after reviewing risk reports, we did not see detailed deliberation in the baseline condition. By providing tangible examples of harm, RiskRAG encouraged more thoughtful, self-reflective decision-making, shifting model selection from a purely technical focus to a more cautious, risk-aware approach.

2. Effective risk communication: Participants valued consistent, comprehensive, actionable insights about potential risks. "*It has a much more detailed and comprehensive analysis and also a balanced assessment.*" (P5, med). "*I choose Risk Heatmap because it is a valuable tool for identifying, assessing, and communicating risk visually and effectively with broad contexts.*" (P2, dev).

3. Accessible presentation: This theme emphasizes the role of visuals and accessible formatting in usability. "*I prefer risk report A [RiskRAG] because it is visually appealing, elaborate, and uses simple and understandable language. Risk Report B is shallow and plain, and does not give people the motivation to read and comprehend what it reports.*" (P28, med). Even developers, the most knowledgeable cohort, acknowledged that the report simplified technical jargon, making it easier to understand. "*The risks were clearly outlined... which helps demystify the technical jargon and allows users to make ethical decisions.*" (P30, dev).

4. Risk prioritization for mitigations: Participants noted that the structured categorization supported better prioritization and planning. "*I prefer risk report A because the heatmap type made it easier to spot key risks quickly and prioritize actions.*" (P36, des).

5. Initial learning curve: A challenge emerged, with some participants citing the initial cognitive load required to familiarize themselves with the report. "*Though understanding the heatmap was a bit of a struggle, it gets easier once you understand.*" (P1, dev).

For the users who preferred the baseline, two key themes emerged, highlighting potential areas for improvement for RiskRAG.

1. Detailed textual explanations: Baseline allowed easier interpretation for some participants. "*The text-based format allows for more comprehensive explanations of the risks and mitigations.*" (P34, des). *"The structured text clarified the nuances and provided specific examples, making it easier to evaluate the ethical implications.*" (P22, dev).

2. Familiarity and experience: Participants highlighted the comfort they felt with the familiar textual style of baseline risk content, aligning with their prior experiences in similar tasks. "*Because it was something I had encountered before this study... made it easy for me to express my thoughts.*" (P11, dev).

6.5.4 Ethics

All three studies, including co-design and evaluation, were approved by the authors' organization. Participants received informed consent detailing the study's purpose, data usage,

Table 6.7: Comparison of RiskRAG to closely related prior works in two main research areas: AI risk documentation, and tools for populating such documentation. For the AI risk documentation solutions, we examine whether they address each of our identified design requirements and whether they are specifically designed for model cards. For tools populating the AI risk documentation, we assess whether their content meets the requirements, is tailored to model cards, and whether it leverages RAG techniques. AI documentation proposals, Risk Cards, and Model Cards in 2024 lack prioritization of risks (R5), do not address mitigation strategies (R4), or structure them according to taxonomies (R2). Tools for populating AI risk documentation, such as ExploreGen, AHA!, FarSight, and CardGen do not focus on generating model-specific risks (R1), mitigation strategies (R4), or to prioritize them (R5), and they mainly rely on LLMs only. While CardGen uses RAG techniques, it produces risk content mimicking existing model cards, which itself falls short of the design requirements.

	R1	R2	R3	R4	R5	Designed for model cards	Uses RAG
AI Risk Documentation							
Risk Cards [86]	✗	✓	✓	✗	✗	✗	-
Model Cards 2024 [161]	✓	✗	✓	✓	✗	✓	-
Tools for Populating AI Risk Documentation							
ExploreGen [137]	✗	✗	✓	✗	✗	✗	✗
AHA! [49]	✗	✓	✓	✗	✗	✗	✗
Farsight [324]	✗	✓	✓	✗	✓	✗	✗
CardGen [189]	✓	✗	✗	✗	✗	✓	✓
RiskRAG (ours)	✓	✓	✓	✓	✓	✓	✓

and their rights. Confidentiality and anonymity were ensured, with data handled according to established ethical guidelines. Participants recruited via Prolific were compensated at a minimum rate of $12/hour. Research materials, such as study artifacts and thematic analysis codebook, were shared in compliance with transparency criteria outlined by Salehzadeh Niksirat et al. [274].

6.6 Discussion

In this study, we introduced RiskRAG, a Retrieval Augmented Generation-based system designed to improve the risk reporting process for AI models. Our work addresses a gap in current AI model documentation practices, particularly in model cards, which often lack comprehensive and specific risk assessments tailored to both the model and AI uses. Through iterative co-design and user studies with a total of 181 AI developers, UX designers, and media professionals, we empirically demonstrated that RiskRAG provides more contextualized and actionable risk reports, compared to the baseline model card risk sections. RiskRAG encouraged a more cautious and deliberative approach to model selection, effectively supporting decision-making.

6.6.1 Data-driven Solution for Risk Documentation for AI developers

In contrast to previous efforts that leverage LLMs to envision AI risks [49, 137, 324], RiskRAG is grounded in real-world datasets, ensuring a robust and adaptable solution. It employs retrieval-augmented generation to source human-written risks from model cards or documented real-world harms, minimizing hallucinated or generic risks. RiskRAG enhances existing solutions that replicate predefined formats or focus on subsets of model cards (e.g., CardGen tied to research papers or GitHub repositories [189]). Instead, it works on all model cards, addressing critical gaps in current methodologies for risk reporting as highlighted in Table 6.7. Risk Cards [86] do not include model-specific risks or mitigation strategies, Kennedy-Mayo and Gord [161] omit risk categorization, and neither documentation solutions address risk prioritization. Current solutions offer partial solutions but fail to meet the comprehensive requirements of risk documentation identified. ExploreGen [137] generates uses but not risks, while AHA! [49], and FarSight [324] produce risks for abstract AI systems without tailoring them to specific models or prioritizing them. Additionally, none propose actionable mitigation strategies. In contrast, RiskRAG uniquely prioritizes risks based on real-world harms, while associating each risk with tailored mitigation strategies. However, we believe these previous solutions can complement RiskRAG. For instance, RiskRAG did not show statistical significance in participant responses regarding whether risks were adequately contextualized to the use case or not. Incorporating tools like FarSight, which generate risks for specific applications, could enhance RiskRAG's ability to provide a more comprehensive and context-sensitive risk assessment.

RiskRAG demonstrates strong potential for *generalizability*, offering two key benefits for unseen and lesser-known models: its structured template prompts developers to generate meaningful risk content, and the generated reports serve as a valuable starting point compared to a blank slate. To assess this capability, we selected four lesser-known models from the 450K snapshot of models from HuggingFace (selection process and the models described in Appendix C.5). Manual evaluation of their RiskRAG-generated reports confirmed the utility and relevance of the outputs. These generated reports, shared in Supp. Mat. 3.2, illustrate how RiskRAG can support effective risk documentation even for models lacking extensive prior information, reinforcing its adaptability and broader applicability.

Further, RiskRAG's architecture is designed for *scalability*, enabling seamless integration with larger and more comprehensive datasets, such as the MIT AI Risk Repository[69] [290], the Automation Incident Repository (AIAAIC) Repository[70], and the OECD AI Incident Monitoring system (AIM)[71]. This ensures that RiskRAG remains flexible and future-proof, evolving as new data emerges. Additionally, RiskRAG reports offer clear and actionable mitigation strategies for each identified risk, empowering model users to address potential issues proactively before deploying them. Importantly, we envision RiskRAG not as a final solution but as a tool to support AI developers in producing effective risk reports. Starting with its generated content and structured format, developers can update, omit irrelevant

[69] https://airisk.mit.edu/
[70] https://www.aiaaic.org/aiaaic-repository
[71] https://oecd.ai/en/incidents

risks, and be inspired to produce new ones, fostering a collaborative and iterative approach to AI risk documentation. This process not only streamlines risk assessment but also addresses the challenge of limited motivation for AI developers to identify potential harms [200, 201] by providing a structured, accessible foundation for risk documentation.

6.6.2 Raising Awareness and Promoting Responsible AI Use

As of July 2024, out of 450K model cards on HuggingFace, only 64K included risk-related sections, and just 2672 of those were unique. This means that approximately 86% of model cards on HuggingFace do not mention any risks. These numbers are consistent with findings from previous studies [33, 184]. Our co-design study supports these results, as AI developers reported that they typically focus on the technical aspects of model documentation, often overlooking risk-related sections. However, they also reported feeling enlightened and inspired by the risk content we provided during the study, in alignment with research showing that even AI practitioners and researchers find it challenging to anticipate the risks associated with AI systems and models [44, 98].

The practical potential of RiskRAG lies in its integration with platforms such as HuggingFace, GitHub, Model Zoo[72], PyTorch Hub[73], or Google AI Hub[74]. It can serve as a template or an interactive interface for model card creation, enabling AI developers to prioritize model-specific risks and mitigations contextualized to diverse uses. Through an interactive process, developers can refine risk assessments, retrieve relevant examples of AI incidents, and identify mitigation strategies with reduced effort. Crowdsourced feedback could further enhance RiskRAG, refining its prioritization techniques and producing tailored, actionable reports over time.

Our findings highlight RiskRAG's ability to foster deliberative and cautious decision-making. The preliminary study confirmed that RiskRAG met all desired requirements, while the final study demonstrated its effectiveness in enhancing users' ability, developers, designers, and media professionals alike, to identify risks, devise mitigations, and improve explanation quality. By deepening users' understanding of model impacts, RiskRAG can enable more informed decision-making, such as opting against unsuitable models, strengthening risk management, and making critical adaptations prior to deployment. This approach helps prevent harmful incidents and promotes ethical AI use, ensuring AI technologies are developed and deployed in alignment with responsible and safe practices.

6.6.3 Broader Implications: Model Cards, Public Outreach, and Policy Making

Recent efforts have sought to enhance the ethical considerations and risk sections in model cards [86, 161]. The current standard for documenting AI models—model cards—can be significantly enhanced through the integration of insights from our work with RiskRAG. Specifically, the risk sections could be transformed to reflect the detailed risk assessments

[72] `https://modelzoo.co/`
[73] https://pytorch.org/hub/
[74] `https://cloud.google.com/vertex-ai/`

generated by RiskRAG, making risk documentation more comprehensive. This enhancement has the potential to establish a new best practice, encouraging deeper engagement with potential risks. Consequently, this could position model cards as robust foundations for impact assessment reports [37, 38, 203], enabling thorough evaluations before deployment. RiskRAG also opens new research avenues regarding the presentation of use-specific risks and mitigations in model documentation. By tailoring risk information to specific use cases, model cards could evolve to resemble impact assessment reports, offering a more structured approach to decision-making about the consequences of deploying AI models in various contexts.

Beyond risk sections, our user-centred approach to developing RiskRAG could inspire improvements across other sections of model cards, such as model and data specifications. By optimizing the presentation and usability of model cards, they can become more accessible, informative, and effective for developers at all levels.

As regulatory scrutiny [73, 145] around AI technologies increases, businesses and developers must ensure compliance with evolving laws. RiskRAG can streamline this process by identifying risks and aligning its reporting with specific regulatory requirements, reducing the complexity of navigating legal frameworks, and helping organizations address potential compliance gaps proactively. This alignment ensures models meet regulatory standards from the outset. Moreover, RiskRAG has the potential to enhance the accessibility of risk information for both the *general public* and *policymakers*. Participants in our study highlighted concerns about the lack of transparency in traditional risk reports, perceiving missing or unclear information as intentional, affecting trust and accountability. As one participant noted, *"It seems in their documentation to pretend like they're conveying actual information."* (P25, dev), while another stated, *"It made me feel as though it was actually hiding information about its risks that it would rather have people not know."* (P24, dev). By presenting clearer and more comprehensive insights, RiskRAG could foster greater trust, enabling stakeholders to make *better-informed* decisions and promoting more responsible AI governance.

6.6.4 Limitations

RAG Model Biases and Hallucinations. Retrieval-Augmented Generation, while better than Large Language Models at grounding generated information in external databases and reducing hallucinations, has limitations. Retrieval bias may occur if certain model cards are over-represented, though our user studies with diverse model types (e.g., text generation, image and text-to-text, speech-to-text) showed high-quality risk generation. By incorporating real-world incident risks, we further reduced reliance on model cards and captured undocumented risks. Frequency bias in risk prioritization, based on incident frequency, could lead to an over-focus on certain risks and an under-focus on others, if the incident repository, in this case AIID, is biased. This can be mitigated by integrating multiple repositories such as the OECD AIM [236] and the AIAAIC. Despite fewer hallucinations than LLMs [116, 240], RAG systems still occasionally generate incorrect risks. For example, RiskRAG incorrectly flagged "generates disinformation by creating misleading or false images" for a model that processes but does not generate images (i.e., an image-to-text model). In our experiments,

emphasizing model type, as well as types of input and output of the model in prompts, minimized such errors. Since RiskRAG is designed to assist, not replace, developers, these errors pose limited risk, as developers can and are expected to refine outputs. Future work could add a secondary generative agent to verify risks and further reduce hallucinations.

Lastly, we acknowledge that some risks extracted by our method may not be entirely relevant to a specific model. In our evaluation detailed in §6.5.1, we automatically validated the quality and relevance of the retrieved risks; however, due to the challenge of recruiting highly knowledgeable AI and ethical experts for specific models, we could not conduct a large-scale expert study on this issue. Instead, we manually assessed the quality of risks for two sets of models: four models for which the three authors had high expertise, and four less-known models (discussed in Appendix C.5). When asked, *"Are the risks relevant to the model and its type?"* on a scale from 1 to 5, the average response across all models was 4.75, indicating strong agreement on their relevance.

Limitations of Evaluation Data. We selected the most popular model cards for our pseudo ground truth dataset, which may inadvertently favor newer models that are closely aligned with existing and well-known models. To verify RiskRAG's performance on lesser-known models, we conducted an additional experiment (Appendix C.5), which confirmed its effectiveness. Future work could explore incorporating additional information specific to the new models (e.g., their associated paper, similar to CardGen) to enhance output quality in such cases.

Artificial Setting and Study Scope. Although we worked with 165 AI stakeholders across two studies and provided realistic tasks, the studies were conducted in controlled settings rather than real-world environments, and they were one-time experiments rather than longitudinal. This limits the external validity of our findings, especially concerning RiskRAG's long-term performance.

Learning Curve and Adapting to RiskRAG. Thematic analysis revealed that some participants who favored RiskRAG initially faced challenges in adapting to it, while those who preferred baseline reports often cited familiarity as the key factor for their choice. This disparity, alongside preference scores from the final study—58% for developers compared to 70% for media professionals—suggests that AI stakeholders accustomed to traditional reports may require additional effort to adjust to RiskRAG. Future research should explore a hybrid approach, blending RiskRAG's visually structured matrix with textual explanations and tabular examples common in existing reports, as proposed by some of our participants: *"Although the heatmap allows you to see at a glance, I think I prefer the text... I think the ideal would be a combination of both"* (P6, des).

Incomplete (Systemic) Risk Coverage. Although participants perceived RiskRAG's reports to be more detailed, there is a chance that risks derived from similar models could still be incorrect, as discussed above. Furthermore, systemic risks, which take longer to materialize, are difficult to fully capture even with real-world incident data [317]. Thus, RiskRAG's coverage of systemic harms may still be incomplete.

Pretrained Models. We used pretrained embedding models for RiskRAG without further fine-tuning, based on prior work suggesting this approach generally works well. However, given the complexity of generating detailed risk content, future research should explore whether *pretraining RAG components* on model card data or real-world incidents could

improve the quality of risk assessments.

Static Presentation. We focused on static PDF presentations of model cards, but it is likely that an interactive format would be even more effective. Future work should explore how developers interact with dynamic, interactive versions of RiskRAG, where risks and mitigations adapt based on the selected use case. Additionally, a community-driven feedback mechanism could invite developers to report new risks encountered during production, enriching the risk database over time.

6.7 Conclusion

Our work on RiskRAG demonstrates the potential of a data-driven, AI-assisted risk reporting system that aligns with the needs of AI developers. By addressing gaps in current model card risk reporting and providing actionable insights, RiskRAG fosters responsible AI use and improves risk documentation practices. With further refinement and adoption, RiskRAG could significantly enhance how AI models are evaluated and deployed in the real world, contributing to safer and more transparent AI systems.

Chapter 7

Discussion

This discussion chapter moves beyond the findings of the five individual studies to synthesize them into a broader narrative. The aim here is to draw connections across them, identifying common patterns and cumulative insights. The discussion proceeds in several parts. First, it summarizes the contributions of this thesis in light of the studies conducted. Then, it synthesizes the key findings across the five studies, emphasizing the ways they complement and extend one another. Following this, it turns to the practical and policy implications of the work for developers, job seekers, policymakers, and regulators. Finally, the chapter considers the limitations of this body of work and ends by proposing directions for future research that extend beyond the scope of the present studies.

Through this integrative discussion, the chapter argues that addressing risks and fairness in AI requires not just better models, but better infrastructures, design processes, and collective imagination. Responsible AI should be treated as an evolving practice, not as an end product.

7.1 Summary of Contributions

This thesis investigates how to design and evaluate AI-assisted interview training systems and develops approaches for identifying and reporting risks of such AI systems. It began with questions about the role of interview training in fostering motivation and performance (RQ1), the needs of job seekers for responsible AI design (RQ2), bias in AI hiring evaluations (RQ3), and effective methods of reporting risks of AI (RQ4). The preceding chapters sought to address these questions. Across five interlinked chapters, it follows a deliberate progression: from human-centered training design, to stakeholder-driven requirements and risks, to empirical evidence of cultural bias, and finally to infrastructures for risk identification and reporting. Together, these studies contribute to a deeper understanding of what it means to build responsible AI systems that function well and also align with the needs, values, and safety of those they impact.

Chapter 2 introduced the first *artifact* contribution of the thesis: **an asynchronous video interview training tool named InterviewApp**, designed to foster motivation and improve interview performance by supporting the three basic psychological needs: autonomy, competence, and relatedness, as defined by Self-Determination Theory (SDT). The

four-month evaluation study with job seekers demonstrated that InterviewApp had a significant influence on autonomy. The self-determined action of requesting feedback positively impacted both autonomy and relatedness. Engagement with recording and feedback features led to measurable improvements in performance. Notably, the effect of training on performance was mediated through autonomy, but this association between autonomy and performance was unexpectedly negative.

Chapter 2 discusses this result extensively: for many job seekers, extrinsic pressures, such as the urgent need to secure employment, could have fuelled greater engagement and performance, especially in the short term. These **external motivators, while potentially crowding out intrinsic motivation and explaining a decrease in autonomy, may be essential to sustaining early involvement in training**. However, to sustain motivation to train in the long term to develop inherent skills beyond the immediate goal of securing the next job, autonomy becomes critical. Over time, autonomy-supportive tools like InterviewApp could enable job seekers to transition into more self-directed autonomous learning, helping them develop transferable skills that support long-term growth and well-being. The findings contribute to the *empirical* evidence that **autonomy plays a key mediating role in influencing motivation and interview performance**. The study also highlighted key design challenges, including the limited effectiveness of peer-based feedback and the absence of mechanisms to support long-term scaffolding without expert feedback. These limitations underscore the need for more scalable, adaptive forms of support.

Drawing on insights from Chapter 2, where job seekers anticipated the use of AI feedback due to the lack of peer and expert reviews, Chapter 3 presents the next major contribution of this thesis: an *empirical* investigation into the participatory **design of AI-assisted interview training systems through a multi-phase co-design methodology**. This study involved both job seekers and AI practitioners to elicit design requirements, feasibility constraints, and perceptions of responsible AI design and development.

The progressive co-design sessions surfaced a wide range of requirements, including anthropomorphism, personalization, actionable feedback, data diversity, and transparency. While both stakeholder groups broadly aligned on these priorities, notable divergences emerged around accountability, human-AI collaboration, and long-term risks. In particular, job seekers emphasized the need for human oversight, contestability, and ability to correct AI-generated outputs (e.g., transcripts or feedback), concerns that only emerged in practitioner sessions after integrating user perspectives. This chapter also identified four perceived long-term risks of AI-based training raised by job seekers: **homogenization of communication, overreliance on AI, illusion of human interaction, and reinforcement of stress**. These findings contribute to empirical **insights on the design requirements and user-perceived risks for AI-assisted interview training**.

The progressive and sequential structure of the study enabled design iteration across groups. This helped promote a more equitable distribution of decision-making power between users and practitioners in design and underscored the value and the challenges of participatory approaches in designing AI for socially sensitive domains. The multi-phase co-design approach serves as formative groundwork, offering **methodological insights for participatory design in high-risk AI contexts** and warranting future validation. These findings underscore how indirect, phased engagement can surface overlooked issues and foster

more inclusive, accountable AI system design.

Chapter 4 addresses a key concern raised by job seekers in Chapter 3: the fairness of LLM-based evaluations in interview assessment. This study conducted an *empirical* evaluation of interview transcripts from Indian and UK candidates, using two widely used large language models (GPT-4o and Gemini Flash). The transcripts were scored on key job attributes such as hireability, positive impression, self-promotion, and storytelling. **Indian candidates received significantly lower evaluations across job attributes** even when identity-related entities were anonymized, suggesting that communication styles played a critical role. Linguistic analyses revealed that features such as type-to-token ratio, sentence length, and readability significantly influenced LLM scoring. However, the candidate region (UK vs. India) remained a significant predictor of scores even after controlling for these variables, indicating the presence of embedded cultural preferences in the models.

This chapter contributes **empirical evidence of cultural disparities in AI-based interview evaluations and highlights the risk of homogenization**, where candidates may feel pressured to adopt dominant linguistic norms to perform well. These findings challenge assumptions of AI neutrality and underscore the importance of culturally aware design and evaluation practices in high-risk applications.

The cumulative insights from Chapter 3 and Chapter 4 point to a broader issue of risk identification and reporting for AI models in high-risk domains like employment. Chapter 5 examined the current state of AI model risk documentation, aiming to understand how developers reported risks associated with different AI models. By analyzing nearly 450,000 model cards from HuggingFace, a widely used model hosting platform, the study found that only around 64,000 included risk-related sections, revealing that risk reporting is largely absent. From this subset, 2,863 unique risks were extracted and validated to construct the **AI Model Risk Catalog, a structured dataset linking categorized risks to specific models**.

To understand where AI developers' risk reporting aligns (or not) with other perspectives, the chapter compared three sources: developer-reported risks from model cards, researcher-identified risks from the MIT Risk Repository, and real-world harms from the AI Incident Database. Developers most often reported technical issues like model limitations, safety, and bias, which aligned with over half of the documented incidents. In contrast, researchers emphasized societal, governance, and agency-related risks, which were far less represented in actual incidents. However, **the largest share of real-world incidents involves malicious use and misinformation risk types underrepresented by both developers and researchers**. This chapter provides a dataset contribution and an *empirical* basis for understanding how developers report, define, and document model-specific AI risks, thereby complementing existing efforts to map, manage, and audit them.

To address the challenges of AI model risk reporting identified in Chapter 5, the final Chapter 6 introduces **RiskRAG, a retrieval-augmented generation system designed to support developers in creating improved risk reports**. With insights drawn from literature and co-designed with AI developers, RiskRAG is guided by five design requirements: it is built to identify diverse model-specific risks, present them in a clear way, prioritize risks, contextualize them for real-world use, and propose actionable mitigation strategies. Grounded in human-written risks from model cards and AI incident databases, RiskRAG

produces reports that align with these design goals.

Together, the RiskRAG system and its enhanced risk reports constitute the second *artifact* contribution of the thesis. A baseline evaluation showed that RiskRAG successfully retrieved most of the risks already present in high-quality model cards. A preliminary study with AI developers found that RiskRAG was preferred over standard model cards for its ability to meet all five design requirements. Further evaluation with developers, designers, and users provided *empirical* evidence that **RiskRAG reports improve clarity, encourage more cautious decision-making, and scaffold developers' ability to reflect on potential harms**. The system offers a scalable and practical approach to making risk documentation more actionable, contextual, and trustworthy.

7.2 Synthesis

This thesis was structured around two interdependent areas of inquiry: (*i*) human-centered design and evaluation of AI-assisted interview training systems, and (*ii*) responsible identification and reporting of risks in high-risk AI systems. Taken together, the empirical findings contribute to both areas in interconnected ways.

Chapter 2 demonstrated how motivational needs such as autonomy, competence, and relatedness shape engagement with interview training tools. It particularly highlighted the mediating role of autonomy in influencing performance. The findings enhanced the fundamental understanding of designing interview training systems to support motivation and how job seekers interact with them, impacting their performance (area *i*).

Chapter 3 contributes to both areas: it highlighted job seekers' needs and functionalities that inherently supported autonomy, competence, and relatedness (area *i*), while simultaneously emphasizing responsible developer practices and surfacing socio-psychological risks such as homogenization and overreliance that rarely appeared in developer-led risk reporting (area *ii*).

Chapter 4 further bridges both areas: it demonstrated how human-centered evaluation (area *i*) of LLM hiring assessments surfaced cultural biases by penalizing certain groups of people, their linguistic markers, and styles of communication (area *ii*). On one hand, it demonstrated how LLM-based evaluations threaten autonomy and competence by penalizing diverse communication styles (area *i*); on the other, it showed that these disparities demand stronger, use-specific risk documentation practices (area *ii*).

Finally, Chapters 5 and 6 deepen the risk-reporting focus. Chapter 5 examined the current risk reporting status in model cards, highlighting the categories of risks over- and under-reported by developers compared to researchers and real-world incidents (area *ii*). Chapter 5 proposed a solution for enhanced risk reporting (area *ii*) guided by literature and co-designed with developers. These extend earlier insights by showing that developer-led documentation practices rarely capture the experiential harms identified earlier, underscoring the need for risk infrastructures that integrate user-centered perspectives.

This trajectory shows that the two areas of inquiry are interdependent. Insights from human-centered design of training systems reveal classes of potential harms that must be treated as risks, while shortcomings in risk documentation explain why these harms remain

overlooked and unaddressed. The following subsections explain these insights that emerge when the chapters are examined collectively under a single lens.

7.2.1 Motivation and Autonomy Support in Training

Chapter 2 demonstrated that autonomy plays a critical role in shaping job seekers' engagement with interview training using InterviewApp. The intervention had a significant impact on job seekers' autonomy. The ability to request feedback, an act of self-determined choice, positively influenced users' sense of autonomy. Autonomy also mediated the relationship between training and performance. In other words, the size of the effect of training on performance could be explained by the autonomy possessed by the job seeker. However, this relationship was unexpectedly negative: job seekers with stronger extrinsic motivation (e.g., those urgently seeking employment or due to social pressure) showed higher engagement and better performance in the short term compared to those with more intrinsic, long-term aspirations.

While Self-Determination Theory positions autonomy as central to sustained motivation, these findings reveal a more nuanced and complicated view of autonomy, explaining the spectrum of motivation (refer to Sections 1.2 and 2.2) in the context of interview training. Autonomy alone may not be sufficient, especially in high-pressure, resource-constrained contexts like job search. External motivators may be necessary for some individuals to initiate and sustain effort until intrinsic interest or skill mastery develops [59]. This points to the need for training tools that support both immediate performance goals and long-term growth.

Autonomy becomes especially valuable when training systems are designed to help job seekers accrue transferable skills, fostering self-directed learning that extends beyond a single job search. In Chapter 3, this theme reemerges through **job seekers' desire and emphasis on design features that support autonomy, competence, and relatedness: personalization, flexibility, transparency, and human-like interactions.** Requests for explainability and system accountability reflected not just a desire to understand AI decisions, but a need for meaningful control over the training experience. Providing clear rationales and the ability to correct or influence system outputs fosters volitional engagement and deepens trust [266]. Moreover, **collaborative interactions where users help improve the system, such as by correcting AI misinterpretations, have been shown to reduce the trust gap [154, 164] and enhance both autonomy and well-being** [204, 331, 353]. When users perceive that their input shapes the system, they are more likely to experience a sense of agency, motivation, and psychological investment in the training process.

Further, job seekers' actions to request feedback had an impact on their relatedness, reflecting attempts to meet their need for relatedness. In Chapters 2 and 3, participants expressed a desire for more human-like qualities in training systems, including conversational interfaces, anthropomorphic elements, and empathetic feedback. Combined with the findings from Chapter 2, where participants lacked a supportive peer network, these early indications suggest that job seekers might be seeking alternative forms of social connection through AI. While these findings are not conclusive, they raise important questions about whether and

how AI systems may begin to serve as relatedness support, an area that warrants deeper investigation.

Across both chapters, a common insight emerges: AI-based training systems should not merely optimize for efficiency or assessment accuracy, but must be designed to support motivational sustainability, enabling engagement across diverse user needs, life circumstances, and long-term career trajectories through autonomy-enhancing and adaptive design.

7.2.2 Contextualizing Fairness Risks to Uses

Chapter 4 demonstrated that cultural bias in LLM-based evaluations persists even under controlled, anonymized conditions. Indian candidates consistently received lower scores than their UK counterparts across key job-relevant attributes, despite the removal of identity markers. These disparities point to deeply embedded communicative norms and implicit preferences within the model, biases that are not easily eliminated through technical fixes or anonymization.

This persistence of cultural bias becomes even more evident when viewed alongside the findings in Chapter 5. The AI Model Risk Catalog shows that developers do report fairness-related risks: 24% of reported risks fall under *unfair discrimination and misrepresentation*, and another 8% under *unequal performance across groups*. These categories represent only 6% and 2%, respectively, in the MIT Risk Repository, where researchers focus more on systemic and governance-related risks. Yet, as Chapter 4 highlights, bias is not *one-size-fits-all*. **Bias, risks, and limitations manifest differently depending on the application, particularly in high-risk domains** like employment, where cultural mismatch in language or tone can have direct consequences for inclusion and opportunity. This underscores a key limitation in current risk documentation practices: simply naming "bias" in general terms is insufficient. Reporting risks in broad terms leads to vague concerns, which can obscure real vulnerabilities.

Effective mitigation requires that **risks be documented in direct relation to the model's intended uses, rather than at a generic level**. This need was explicitly raised by developers during the co-design session for improved risk reporting in Chapter 6, who emphasized that meaningful risk documentation must reflect how models are applied in specific contexts. RiskRAG is a step toward such use-aware documentation. By retrieving and generating risks and contextualizing them to specific applications, RiskRAG pushes developers to consider not just whether bias might occur, but how, where, and for whom. This orientation aligns with emerging tools such as use case cards [148] and AI cards [125], inspired by regulatory frameworks like the EU AI Act [126], which emphasize use-specific risk transparency and accountability.

7.2.3 Participatory Risk Infrastructures

A third cross-cutting theme emerging from this thesis is the gap between documented risks and those experienced by end users, particularly in the form of socio-psychological harms like *homogenization*. In the co-design study (Chapter 3), job seekers voiced concern that AI-based interview training systems subtly pressure them to conform to standardized, "ideal" styles

Table 7.1: Socio-psychological risks raised by job seekers, mapped to SDT needs and aligned with risk sub-categories from the MIT AI Risk Repository [290] and the DeepMind taxonomy of harms [329]. Across both frameworks, these risks primarily fall under the Human–Computer Interaction (MIT) and Human Autonomy & Integrity (DeepMind) category, and the corresponding sub-categories are shown in the table.

Risks raised by job seekers (sub-categories)	**SDT Need Impacted**	**Explanation**
Homogenization of communication styles DeepMind: Persuasion & Manipulation MIT: Loss of human agency and autonomy	Autonomy *(primary)*; Competence	AI nudges job seekers toward narrow, "ideal" ways of speaking, suppressing cultural and linguistic individuality. This forces conformity (autonomy) and narrows opportunities for skill and mastery (competence).
Overreliance on AI judgments DeepMind: Overreliance MIT: Overreliance and unsafe use	Autonomy *(primary)*; Competence	Job seekers rely excessively on AI feedback and scores, replacing self-judgement with system guidance (autonomy) and risking decline in skills or miscalibration (competence).
Illusion of human interaction & misalignment with real hiring DeepMind: Persuasion & Manipulation (*primary*), Overreliance MIT: Overreliance and unsafe use (*primary*), Loss of agency	Relatedness *(primary)*; Autonomy; Competence	Job seekers treat AI as a recruiter, creating simulated connections (relatedness), steering choices through persuasive design (autonomy), and misaligning preparation compared to real recruiter expectations (competence).
Reinforcement of insecurity and stress DeepMind: Violation of personal integrity MIT: Loss of human agency and autonomy	Competence *(primary)*; Autonomy	Negative AI feedback overwhelms and reduces perceived efficacy (competence), erodes control over learning (autonomy), leading to heightened stress and alienation.

of communication, eroding individuality and reinforcing stress. This concern was empirically supported in Chapter 4, where LLM-generated scores consistently favored UK-based transcripts over Indian ones, even after anonymization. Together, these findings suggest that cultural mismatches embedded in model behavior can systemically undervalue diverse communication styles and reproduce narrow normative expectations.

When viewed through the lens of SDT, the **socio-psychological risks identified by job seekers map directly onto the basic needs of autonomy, competence, and relatedness.** To situate them better, Table 7.1 maps job seeker-elicited risks to categories in the MIT AI Risk Repository taxonomy [290] and the DeepMind taxonomy of harms [329] as well as to the three basic psychological needs from SDT. These risks primarily fall under *human–computer interaction* (MIT) and *human autonomy & integrity* (DeepMind) areas. This underscores that these **risks are primarily user-facing, stemming from AI-job seeker interaction dynamics** rather than core technical failures. For these risks, MIT taxonomy categorizes them as loss of agency and overreliance, while DeepMind adds more granular sub-areas, highlighting persuasion, manipulation, and personal integrity. Table 7.1 also makes clear that the risks identified by job seekers are directly tied to the needs: homogenization and overreliance erode autonomy, reinforcement of stress and insecurity due to judgmental feedback or pressure to perform undermines competence, and the illusion of human interaction affects relatedness. In other words, what users describe as experiential harms align closely with both formal risk categories of human-computer interaction and core psychological needs.

However, **these risk categories are rarely emphasized in developer-reported risks.** The AI Model Risk Catalog (Chapter 5 and Figure B.3) reveals that only 4% of developer-reported risks fall under the *human autonomy & integrity*: 1% mention persuasion and manipulation, 1.2% mention overreliance, and just 0.6% cite violations of personal integrity. Only overreliance on AI appears in both developer and user-elicited concerns, whereas the other three user-elicited risks do not appear in developer-reported risks. This asymmetry in risk identification reflects a critical insight: while developers and researchers tend to focus on measurable technical limitations, abstract fairness, or broad social impact, users are concerned about *experiential harms* that thwart their basic needs, leading to frustration. These arise from interacting with AI systems in emotionally charged, high-risk contexts like hiring. Notably, the majority of real-world incidents occur at this human–AI interaction layer, where individuals engage with AI systems directly [317].

Chapters 3 and 6 point toward corrective approaches through participatory risk infrastructures. Chapters 3 shows how phased co-design can surface user-facing risks that often remain invisible in developer-led processes. RiskRAG contextualizes risk documentation using structured data and real-world incidents. Importantly, the need to contextualize risks to the model's intended use was explicitly raised by developers during RiskRAG's co-design, recognizing the need to surface risks from end-users.

Broader shifts are needed in how harms are conceptualized, elicited, and valued, particularly those that are cultural, psychological, and longitudinal. **Participatory methods and use-aware documentation are early but essential steps** toward surfacing these overlooked dimensions. Mitigating these requires, including better documentation practices, a rethinking of how we design AI systems. **Self-Determination Theory provides a frame-**

work for such mitigation of socio-psychological risks: supporting autonomy by giving users meaningful choice and control over when and how to rely on AI recommendations; fostering competence through transparent, skill-building feedback; and nurturing relatedness by enabling authentic, constructive forms of connection rather than shallow simulations of social interaction. Designing systems with these needs in mind not only improves user experience and engagement but also reduces the likelihood of socio-psychological risks and promotes well-being, aligning motivational sustainability with responsible risk governance.

7.3 Implications

This thesis offers important takeaways for multiple stakeholder groups involved in the development, use, governance, and deployment of AI systems in high-risk domains like employment and training. Based on the findings from five chapters and their synthesis, this section outlines the practical and policy implications for AI practitioners, end users, governance bodies, and industry stakeholders.

7.3.1 Implications for Job Seekers

This thesis offers several insights into how job seekers might navigate, engage with, and critically assess AI-assisted training and evaluation systems. While these systems promise personalized support and scalability, job seekers should remain attentive to both their empowering and alienating effects.

Balancing external motivators with long-term autonomy. As shown in Chapter 2, job seekers engage with training under varying degrees of perceived autonomy. External motivators, such as the urgent need to secure employment, can initially drive engagement. However, long-term growth depends on cultivating autonomy. Autonomy encourages independence and decision-making and enables job seekers to pursue transferable skills and self-directed learning that extend beyond a single job search. While its benefits may not be immediately visible, maintaining autonomy over time supports deeper learning and long-term career resilience. Job seekers should therefore strike a balance: engaging with short-term demands without losing sight of their long-term agency. This means choosing tools that provide choice, customization, and the ability to question or adapt AI feedback; empowering them to take ownership of their preparation and align it with personal goals.

Transparency, accountability, and actionable feedback were recurring themes in both Chapter 2 and Chapter 3. Job seekers benefit most from tools that clearly explain why a score was given and how to improve. Systems that offer rationale, highlight specific behaviors, and allow room for contesting AI output can help job seekers learn more effectively and feel more in control of their training.

Navigating the emotional impact of AI. As job seekers in the co-design study expressed, interacting with AI training systems can evoke feelings of being judged, misunderstood, or pressured to conform. This emotional weight, especially when job seekers are

already facing employment uncertainty, should not be overlooked. Job seekers may need to actively guard against overreliance and resist internalizing scores as fixed indicators of performance. Instead, they can treat AI feedback as one of many inputs in their broader preparation process. Again, choosing and engaging with need-supporting tools and systems that provide autonomy over their entire learning process can help.

Fostering community and peer support. A key finding from Chapter 2 was the lack of a supportive peer network. Without access to empathetic human feedback, job seekers may turn to AI systems as proxies for peer or mentor interactions. While AI can scaffold learning, it is not a full substitute for human connection [18, 71]. Job seeker should be actively encouraged to seek or build communities by connecting with others at similar stages of their careers to exchange feedback, share experiences, and reduce isolation. Training systems can play a critical role in facilitating such connections, using AI not as a replacement for empathy but as a bridge that helps job seekers find and support each other.

Ultimately, job seekers may benefit from reframing AI training tools not as evaluators but as potential anchors in their employment journey, providing structure, encouragement, and reflection. Human-centered AI design, combined with peer support networks, may help alleviate the concerns raised and build confidence and sustain motivation over time.

7.3.2 Implications for AI Practitioners and Developers

AI developers, designers, and engineers play a central role in shaping how users engage with AI-assisted training and evaluation systems.

Motivational design must be foregrounded. This thesis demonstrates that technical performance alone is insufficient: developers must also consider how systems support users' psychological needs, particularly autonomy, competence, and relatedness. Design features such as personalized and actionable feedback, transparent rationale and explanations for assessments, and the ability to contest or correct outputs are essential not only for user trust, but also for sustained motivation to engage in the activity, mitigate socio-psychological risks, and promote well-being.

Importantly, these features should be designed to offer flexibility rather than rigid guidance. Developers should strike a balance between standardization and individual expression, allowing users to selectively engage with AI feedback while retaining individual agency. In some cases, developers must also critically assess whether AI is appropriate for a given feature at all.

Participatory design offers critical perspectives. The findings underscore the importance of participatory design in AI development, distributing power to end users in defining system goals and features. Co-designing with end users brings forward nuanced concerns, particularly socio-psychological risks like homogenization and manipulation, that often go unrecognized in developer-led processes. Structured stakeholder engagement expands the scope of risk imagination and promotes design decisions that are more ethically grounded and socially responsive.

Operationalizing socio-psychological risks SDT provides validated psychometric instruments for measuring the satisfaction and frustration of autonomy, competence, and relatedness. Developers can draw on them to assess whether their systems support or undermine autonomy, competence, and relatedness. These instruments can be incorporated into user studies, pilot deployments, or post-use surveys to identify whether design choices pressure conformity, erode agency, or hinder competence. At the same time, standardized scales may not capture AI-specific dynamics such as anthropomorphism or overreliance on automated judgments. Developers should therefore consider combining SDT-based assessments with tailored, context-sensitive indicators of interaction risks. Integrating such hybrid measures into evaluation and documentation pipelines makes socio-psychological risks visible, trackable, and actionable alongside traditional performance and fairness metrics.

Contextualizing evaluation and risk documentation for intended uses. Human-centered evaluation methods should complement traditional benchmarks, capturing how diverse users interact with AI systems in realistic settings. Developers should also clearly communicate a model's limitations, biases, and potential harms in ways that are accessible, understandable, and context-sensitive. Developers should anchor risk reports in specific use cases, reflecting how models perform in situated, high-risk applications. Tools like RiskRAG, which generate risk reports contextualizing specific uses based on structured data and real-world incidents, are a promising step in that direction. They can support developers in anticipating risks, improving documentation quality, and making more informed deployment decisions.

Supporting everyday developer risk reporting practices. Finally, there is a clear need for such tools to be integrated into mainstream model hosting platforms like HuggingFace. Over 86% of model cards currently omit any risk information, echoing earlier findings and insights from co-design that developers often prioritize technical details over risk reporting in documentation. Embedding tools like RiskRAG into development workflows can help shift documentation from a compliance burden to a meaningful design practice, enhancing both accountability and model usability.

7.3.3 Implications for Governance and Policy

As regulatory frameworks such as the EU AI Act and the NIST Risk Management Framework mature, the findings of this thesis offer empirical evidence for shaping more grounded, context-aware, and participatory approaches to AI governance. These implications highlight how current documentation practices and risk governance mechanisms can evolve to better serve affected stakeholders and ensure compliance with emerging regulations.

Use-context must inform risk documentation. The results underscore that risks associated with high-risk AI systems cannot be meaningfully assessed in abstract. Developers tend to report risks in generalized or vague terms, often overlooking how harms manifest in specific use cases. This leads to broad concerns that obscure real vulnerabilities. For

instance, while 24% of risks in the catalog fall under *unfair discrimination and misrepresentation*, the actual cultural biases observed in LLM evaluation in Chapter Chapter 4 were more nuanced and context-specific. Risk documentation must therefore reflect situated use contexts, where real-world interactions and cultural mismatches become legible. The RiskRAG system represents a step in this direction by tailoring risks to the intended uses of models, thereby aligning risk documentation with both regulatory expectations and end-user realities.

Structured formats and clearer standards are needed. Our comparison of three risk sources (the Risk Catalog, MIT Repository, and the AI Incident Database) reveals a lack of standardization in risk definitions, formats, and severity descriptions. While repositories often include long-form ethical narratives and incidents link risks to real-world harms, developer documentation remains brief and inconsistent. A unified standard that specifies harm type, context, severity, and likelihood could improve interoperability and make risk documentation more transparent and actionable. Formats like RiskRAG reports, risk cards [86], and use case cards [148] show promise but require further empirical refinement to ensure usability across stakeholder groups.

Participatory risk infrastructures should be incentivized. Effective governance cannot rely solely on top-down auditing or checklists. Structured stakeholder engagement, as demonstrated in the co-design study, is essential to elicit diverse concerns and lived experiences. For instance, structured and phased co-design sessions brought out risks that are culturally or psychologically situated. These methods align with democratic and pluralistic principles and help surface risks that technical evaluations or benchmark metrics miss. Governance frameworks should therefore incentivize participatory risk identification processes as a best practice in responsible AI development.

Investing in tooling and infrastructure for risk reporting. Tools like RiskRAG demonstrate that meaningful, context-aware risk reporting is feasible when developers are supported. RiskRAG improved clarity, transparency, and perceived usefulness of model cards. Moreover, when integrated into model-hosting platforms (e.g., HuggingFace), such tools can scale best practices and reduce compliance burden. Governance agencies should consider funding or mandating tooling support for risk documentation.

RiskRAG also has the potential to bridge gaps between regulatory expectations and public understanding. Participants in our study expressed concern that traditional risk documentation often appeared misleading. By offering clearer and use-specific risk information, tools like RiskRAG can enhance public trust, inform policymaking, and support better impact assessments. Policymakers can also draw on structured repositories like the AI Model Risk Catalog to develop more targeted and evidence-based interventions.

In sum, this thesis argues that responsible AI governance must not only mandate risk documentation but also reimagine the process as participatory, contextual, and infrastructurally supported. Doing so will help ensure that risk reporting meets the needs of developers, regulators, and affected communities alike.

7.3.4 Implications for Risk Documentation Practices

Centering human-AI interaction harms. As seen in Chapters 3 and 4, the most salient risks, such as reinforcement of stress, illusion of human interaction, and cultural homogenization, stem not from technical failures but from how users interact with and are shaped by AI systems. Yet these harms are largely absent from developer-authored documentation in Chapter 5, where only 4% of risks fall under categories like persuasion, overreliance, or violation of personal integrity. 7% of researcher-reported risks fall under human-computer interaction risks. Risk documentation practices must therefore expand beyond algorithmic accuracy or fairness to include socio-psychological and experiential harms.

Toward participatory risk infrastructures. Effective documentation requires stakeholder involvement throughout the design and deployment pipeline. Envisioning risks, especially interaction risks that involve unpredictable human behaviour, is challenging. As shown in Chapters 3 and 6, participatory methods can help surface overlooked harms and align documentation with real-world concerns. Risk reporting tools should include mechanisms for user input and feedback, enabling documentation to evolve with the system and its users.

Balancing structure with contextual flexibility. Documentation formats must evolve to support both standardization and contextualization. On one hand, structured templates like model cards enable comparison, auditability, and scale. On the other hand, risks are deeply situated in use, varying across domains, user populations, and deployment contexts. Future documentation infrastructures must allow flexibility for articulating context-specific risks while retaining enough structure to ensure consistency. For example, RiskRAG demonstrates the first steps in how risk reporting can be anchored in specific applications and supported by real-world incidents and model-specific risks. Future risk documentation should allow flexible extensions or annotations tailored to the model's use case, while retaining a consistent core schema.

Enabling reflective and iterative practices. Documentation should move beyond checklist-style compliance. It should support critical reflection and iterative refinement over time. Developers often skip or underdevelop risk sections mostly because they lack support to surface complex, evolving risks. Risk reporting infrastructures like RiskRAG should include prompts and scaffolds that encourage developers to consider what might be unknown, who might be affected, and under what conditions a model could cause harm. This can foster deeper engagement with responsible development.

Supporting multi-source traceability. Findings across Chapters 5 and 3 revealed stark differences between developer-reported risks, researcher-reported risks, co-design-elicited concerns, and real-world incidents. To address this, documentation practices should explicitly tag risks by source (such as developer, user, incident database, or literature), and specify the context and level of supporting evidence. For instance, the RiskRAG report identifies risks as being sourced from real-world incidents or model cards and prioritizes them accordingly.

This improves transparency, enables triangulation, and allows downstream stakeholders to assess reliability and relevance.

7.3.5 Design Recommendations for Future Training and Evaluation Systems

Designing for long-term motivation and autonomy. Most systems treat training as a one-off event. AI-based training systems must go beyond optimizing short-term performance; they should support users' motivational needs over an extended, uncertain, and emotionally charged employment journey. Designers should embed affordances for autonomy such as personalized learning, transparent rationales, control over learning pace, and the ability to contest or adapt system outputs. Systems must strike a careful balance: offering structured scaffolding to meet urgent extrinsic goals while gradually fostering self-directed learning and long-term skill development. Future systems could support longitudinal reflection by helping users recognize and track transferable skills acquired over multiple job changes, career pivots, or re-entry after breaks. This means surfacing and reinforcing transferable skills like storytelling, communication, or self-promotion. Features such as performance history across mock interviews, customizable scenario-based practice, and archival comparison tools can help job seekers recognize their growth and reuse prior learning. AI-assisted training systems can play a key role in sustaining motivation across discontinuous episodes of employment.

Creating peer-aware and empathetic support ecosystems. Designs must take into account the social isolation often experienced by job seekers, who may lack supportive peer networks. As shown in Chapter 2, this absence of peer support was evident in practice, while Chapter 3 highlighted a desire for systems that feel more empathetic and human-like, perhaps even serving as a stand-in peer. AI training systems should aim to provide both forms of support. First, they should include features that connect users with others at similar stages in their careers, fostering community, shared feedback, and mutual encouragement. Second, systems can incorporate human-like, conversational elements that convey empathy and emotional understanding. While these features can offer valuable support and serve as an emotional anchor during the job search journey, they should complement rather than replace human connections.

Fostering human-AI collaboration and uncertainty-aware design. To foster trust and deep engagement, AI systems should not be treated as opaque or a black box. Instead, systems should promote human-AI collaboration, where users are empowered to question, correct, and contribute to model understanding. This includes expressing uncertainty in outputs, inviting clarification from users when the system is uncertain, and surfacing the rationale behind decisions. Periodic human-in-control oversights, such as expert reviews or peer feedback, agency to contest AI assessments or modify system inputs, can bridge the trust gap and align system assessments with real-world expectations. For instance, the system could offer the option to request an expert review once the job seeker reaches a certain level of proficiency. An expert, equipped with insights from the job seeker's previous

training sessions, can then provide targeted feedback for further improvement. These hybrid mechanisms enhance accountability, trust, and learning outcomes.

Designing culturally aware and regionally fair systems. Biases in model outputs, especially toward non-Western users, emerged clearly in Chapter 4. Dataset diversification is one way to address this, expanding training data to reflect varied dialects, communication styles, and languages. Beyond data diversification, future systems should incorporate region-specific fairness benchmarks (e.g., caste- and religion-related evaluations), evaluation criteria co-designed with local stakeholders, and inclusion of culturally grounded communication styles. AI hiring systems in particular must reflect non-Western norms and avoid mirroring only dominant global standards.

7.3.6 Extending Design Implications Beyond Hiring Contexts

Although this thesis centers on AI-assisted hiring and interview training, the design principles developed here, such as supporting autonomy, competence, and relatedness; enabling contestability; mitigating socio-psychological risks; and grounding evaluation in real contexts, generalize to a wide range of domains where AI systems provide training, feedback, or performance assessment.

Many adjacent fields share similar motivational vulnerabilities and risks of homogenization or overreliance. For instance, on reskilling and upskilling platforms, learners must sustain motivation across long transitions and make sense of evolving, transferable skills in uncertain environments. AI systems in these settings should therefore offer autonomy-supportive scaffolding, longitudinal reflection tools, and culturally sensitive evaluation criteria.

In public speaking and communication coaching, AI-generated feedback interacts directly with users' identity, confidence, and emotional states. Here, actionable yet empathetic feedback, transparency around assessment rationales, and the ability to contest or reinterpret system judgments are essential to prevent conformity pressure or erosion of self-efficacy. Similarly, language learning, creative skill development, and educational tutoring involve iterative practice, subjective performance appraisal, and culturally shaped communication norms.

Across these varied contexts, the core argument remains the same: responsible AI training and evaluation systems must prioritize long-term motivation, cultural and contextual responsiveness, and user agency. The design implications outlined in this thesis, therefore, offer a broader framework for developing human-centered, responsible AI systems that extend beyond recruitment-oriented applications.

7.4 Limitations

While each study in this thesis offers targeted insights into the design, use, and governance of AI-driven interview training and evaluation systems, several overarching limitations merit

discussion. These highlight structural, methodological, and contextual challenges that future work should consider.

Methodological and Experimental Constraints. Many of the studies involved controlled experimental setups rather than real-world deployment. The four-month longitudinal study (Chapter 2) took place in a realistic setting with job seekers who were given the autonomy to practice whenever they wished and as often as they desired. Though more realistic than short-term lab studies, it may still be insufficient to capture the long-term effects of interview training on employment outcomes. Chapter 4 involved job seekers and AI practitioners in separate sessions. Similarly, Chapter 6 evaluated RiskRAG in controlled settings rather than embedded developer workflows. Across studies, these settings may limit validity and obscure emergent behaviors that arise over time.

Participant Demographics and Representation. Several chapters highlight limitations in the demographic makeup of study participants. In Chapter 3, AI practitioners were largely male and from technical backgrounds, while job seekers were predominantly younger and early in their careers. Chapter 4 featured a comparison between UK and Indian job seekers, but the Indian participants were exclusively university students, unlike the more professionally diverse UK group. These mismatches could have introduced potential confounds not captured in the study.

AI tools biases. Many limitations may have stemmed from the AI systems and tools used in the research. Whisper's transcription inaccuracies (noted in Chapter 4) could have introduced error variance across cultures. LLM-based risk extraction (Chapter 5 and Chapter 6) may have involved hallucination and retrieval bias, particularly for under-documented models. Despite mitigation strategies employed like manual validation, multi-source retrieval, and risk standardization, the underlying biases of pretrained models and platform-driven documentation remain a structural limitation.

Lack of ground truths. Some studies do not have established objective ground truths. For example, Chapter 4 had no human evaluator baseline on the interview transcripts, making it difficult to separate genuine performance differences from biased LLM scoring. Similarly, RiskRAG evaluations relied on authors' domain expertise or participant perceptions to assess risk relevance rather than externally verified compliance experts. These constraints reflect broader gaps in the lack of shared evaluation datasets and open ground truth corpora for risk assessment, hiring fairness, particularly across diverse participant backgrounds.

Lack of recruiter involvement. A significant limitation of the thesis is the absence of recruiters and HR specialists, key stakeholders who routinely interpret candidate assessments and make hiring decisions. Their perspectives were not included primarily due to two constraints: (1) organizational access barriers, and (2) the broader scope of the thesis, which focused empirically on job seekers' experiences and technically on AI practitioners' development processes, leaving limited opportunity to conduct additional industry-facing recruitment within project timelines. Nevertheless, recruiters could have provided valuable insights into how interview evaluations are actually interpreted, how organisational norms shape hiring decisions, and how AI-generated feedback or risk information integrates into real recruitment workflows. Their involvement might also have highlighted practical constraints such as time pressure, candidate volume, or legal considerations that influence the feasibility of human-in-control mechanisms and contestability pathways. Future work should

therefore engage recruiters as evaluators, co-designers, and decision-makers to ensure that proposed systems align with the realities of hiring practice and the organisational contexts in which AI tools are deployed.

7.5 Future Work

This thesis opens multiple avenues for future research that can build on the insights generated across the thesis. In closing, I highlight some of the future research directions for the community.

Empirical testing of job seeker-informed design principles. This thesis offers multiple design recommendations based on the needs and preferences of job seekers, such as explainable AI feedback, conversational interfaces, and empathy-driven systems. Their actual impact on performance, engagement, or motivation remains to be rigorously tested. A promising avenue for future work lies in empirically testing how explanation-rich feedback in AI training systems shapes intrinsic motivation. Drawing on Self-Determination Theory and Deci and Ryan [82], I argue in an earlier work [255] that explanations in AI feedback can strengthen the informational aspect of feedback, thus enhancing a user's sense of competence and self-determination. Future experimental studies should compare such explanation-augmented feedback with generic AI scores, assessing outcomes such as motivation, perceived competence, performance improvement, and user trust. This will offer crucial insights into how AI training systems can move from passive assessment toward active, learner-centered coaching.

Exploring AI as a source of social support. Findings in Chapters 2 and 3 suggest that some job seekers may be looking to AI systems to potentially satisfy a need for relatedness in the absence of peer networks. Future work should investigate this emergent behavior further: to what extent can AI simulate social support, and what are the limits and risks of such anthropomorphized interactions? This line of inquiry is especially relevant for socially isolated users navigating challenging life transitions such as unemployment.

Scaffolding developer risk reporting workflows. As we saw in Chapters 5 and 6, there is a growing need to support developers in documenting model risks in ways that are structured, context-aware, and usable across different stakeholder groups. Building on the foundations laid in this thesis, future work should expand RiskRAG beyond evaluation settings into a usable tool for AI developers to actively create or refine their model cards. Such systems could integrate real-world incident data from multiple databases, developer-provided inputs, taxonomies of potential harms, and crowdsourced user-provided feedback to scaffold high-quality model cards or impact reports. An interesting direction would be to empirically investigate how developers interact with such tooling during real-world documentation tasks and examine whether it improves risk awareness, supports reflective practices, reduces reporting burdens, and enhances documentation quality.

Increasing public understanding of AI risks. As seen in Chapter 6, participants expressed frustration with the opacity and vagueness of current risk documentation formats. Future work could explore how to make AI risk reports more accessible and meaningful for broader audiences, including job seekers, journalists, advocacy groups, policymakers, and the general public. This may involve developing tailored visualizations, public-facing summaries, or interactive tools that help non-experts understand the risks associated with specific models or applications. Evaluating how such communication strategies affect trust, engagement, and informed decision-making will be essential to closing the gap between technical risk documentation and public understanding.

7.6 Conclusion

This thesis set out to examine how AI-assisted interview training systems can be designed and evaluated to support diverse users, while also developing responsible approaches for identifying and reporting risks that such AI systems may pose. It did so by raising questions in two interdependent areas of inquiry: (i) the human-centered design and evaluation of AI-assisted interview training systems, and (ii) the responsible identification and reporting of risks in high-risk AI systems. Across five interlinked studies, it foregrounds the perspectives of job seekers and AI practitioners. Drawing from Self-Determination Theory and participatory approaches, it surfaces both the promise and concerns of AI in training and evaluation. The studies revealed that motivation through autonomy, competence, and relatedness is central to how job seekers experience AI training, while at the same time showing that socio-psychological harms tied to these needs and cultural disparities in algorithmic assessment remain systematically under-recognized in current risk documentation frameworks.

By integrating stakeholder needs into the design of AI tools and aligning technical development with human-centered and responsible values, this work offers pathways to build AI systems that are not only effective but also fair and transparent. It also introduces the AI Model Risk Catalog and the RiskRAG system, which can support developers in anticipating and communicating risks more effectively.

Looking ahead, the challenge is not just to make AI systems more intelligent, but to make them more human-centered and responsible. Training systems should go beyond performance metrics, offering not just feedback, but explanation; not just evaluation, but encouragement. Future tools should foster community, reflect diverse cultural norms, and support user autonomy without compromising individuality. Risk documentation must also evolve. It should be participatory, contextualized, and accessible, not only to developers and regulators, but to end users who are most impacted.

Ultimately, this thesis envisions a future where AI systems are co-designed with the people they serve, where motivation and meaning are not sacrificed for optimization, where lived experiences of users are connected with the systemic infrastructures of accountability, and where risk awareness is as integral as accuracy or scale. By embedding values of autonomy and fairness into AI training systems, we can shape technologies that empower rather than merely evaluate and support users not just in getting hired, but in their learning, growth, and well-being.

References

[1] Gavin Abercrombie, Djalel Benbouzid, Paolo Giudici, Delaram Golpayegani, Julio Hernandez, Pierre Noro, Harshvardhan Pandit, Eva Paraschou, Charlie Pownall, Jyoti Prajapati, et al. 2024. A collaborative, Human-Centred taxonomy of AI, algorithmic, and automation harms.

[2] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Large language models associate Muslims with violence. *Nature Machine Intelligence* 3, 6 (2021), 461–463.

[3] Ada Lovelace Institute. 2022. *Algorithmic Impact Assessment: AIA Template*. Ada Lovelace Institute. Retrieved January 22, 2024 from `https://www.adalovelaceinstitute.org/resource/aia-template/`

[4] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. 2025. AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–21. `https://doi.org/10.1145/3706598.3713564`

[5] Alan Agresti. 2011. Score and Pseudo-Score Confidence Intervals for Categorical Data Analysis. *Statistics in Biopharmaceutical Research* 3, 2 (May 2011), 163–172. `https://doi.org/10.1198/sbr.2010.09053`

[6] AI Global: Global AI Incident and Mapping Project. 2025. Where in the World is AI? `https://map.ai-global.org/` online at https://map.ai-global.org/ Accessed: August 2025.

[7] Elham Albaroudi, Taha Mansouri, and Ali Alameer. 2024. A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. *AI* 5, 1 (Feb. 2024), 383–404. `https://doi.org/10.3390/ai5010019`

[8] Elham Albaroudi, Taha Mansouri, and Ali Alameer. 2024. A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. *AI* 5, 1 (2024), 383–404.

[9] Kendra Albert and Maggie Delano. 2021. This Whole Thing Smacks of Gender: Algorithmic Exclusion in Bioimpedance-based Body Composition Analysis. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 342–352. `https://doi.org/10.1145/3442188.3445898`

[10] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. Contestable AI by Design: Towards a Framework. *Minds and Machines* 33, 4 (Dec. 2023), 613–639. https://doi.org/10.1007/s11023-022-09611-z

[11] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.

[12] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making* 20 (2020), 1–9.

[13] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[14] Keith Anderson, Elisabeth André, T. Baur, Sara Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, H. Jones, M. Ochs, C. Pelachaud, Kaśka Porayska-Pomsta, P. Rizzo, and Nicolas Sabouret. 2013. The TARDIS Framework: Intelligent Virtual Agents for Social Coaching in Job Interviews. In *Advances in Computer Entertainment (Lecture Notes in Computer Science)*, Dennis Reidsma, Haruhiro Katayose, and Anton Nijholt (Eds.). Springer International Publishing, Cham, 476–491. https://doi.org/10.1007/978-3-319-03161-3_35

[15] Simran Arora, Avner May, Jian Zhang, and Christopher Ré. 2020. Contextual Embeddings: When Are They Worth It? https://doi.org/10.48550/arXiv.2005.09117 arXiv:2005.09117 [cs]

[16] Arvind Narayanan. 2019. TL;DS - 21 Fairness Definition and Their Politics by Arvind Narayanan. https://shubhamjain0594.github.io/post/tlds-arvind-fairness-definitions/.

[17] Mohammad Atari, Mona Xue, Peter Park, Damián Blasi, and Joseph Henrich. 2023. Which Humans? https://doi.org/10.31234/osf.io/5b26t

[18] David Autor, Caroline Chin, Anna Salomons, and Bryan Seegmiller. 2024. New Frontiers: The Origins and Content of New Work, 1940–2018*. *The Quarterly Journal of Economics* 139, 3 (Aug. 2024), 1399–1465. https://doi.org/10.1093/qje/qjae008

[19] Ricardo Baeza-Yates. 2018. Bias on the Web. *Commun. ACM* 61, 6 (May 2018), 54–61. https://doi.org/10.1145/3209581

[20] Albert Bandura. 1982. Self-Efficacy Mechanism in Human Agency. *American Psychologist* 37 (1982), 122–147. https://doi.org/10.1037/0003-066X.37.2.122

[21] Adrian Bangerter, Paloma Corvalan, and Charlotte Cavin. 2014. Storytelling in the Selection Interview? How Applicants Respond to Past Behavior Questions. *Journal of Business and Psychology* 29, 4 (Dec. 2014), 593–604. https://doi.org/10.1007/s10869-014-9350-0

[22] Laura-Katherine Barker, James W. Moore, D. Joe Olmi, and Kyle Rowsey. 2019. A Comparison of Immediate and Post-Session Feedback with Behavioral Skills Training to Improve Interview Skills in College Students. *Journal of Organizational Behavior Management* 39, 3-4 (Oct. 2019), 145–163. https://doi.org/10.1080/01608061.2019.1632240

[23] Aaron J. Barnes, Yuanyuan Zhang, and Ana Valenzuela. 2024. AI and Culture: Culturally Dependent Responses to AI Systems. *Current Opinion in Psychology* 58 (Aug. 2024), 101838. https://doi.org/10.1016/j.copsyc.2024.101838

[24] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[25] Raiyan Abdul Baten, Famous Clark, and Mohammed Hoque. 2019. Upskilling Together: How Peer-interaction Influences Speaking-skills Development Online. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, New York, NY, USA, 662–668. https://doi.org/10.1109/ACII.2019.8925448

[26] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software* 67, 1 (Oct. 2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[27] Adil Benabou and Fatima Touhami. 2025. Artificial Intelligence in Human Resource Management: A PRISMA-based Systematic Review. *Acta Informatica Pragensia* 14, 3 (Aug. 2025), 489–505. https://doi.org/10.18267/j.aip.264

[28] Emily Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/. *The Gradient* (2019).

[29] Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604.

[30] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[31] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (Sept. 2004), 991–1013. https://doi.org/10.1257/0002828042002561

[32] Hugh Beyer and Karen Holtzblatt. 1999. Contextual Design. *interactions* 6, 1 (Jan. 1999), 32–42. https://doi.org/10.1145/291224.291229

[33] Avinash Bhat, Austin Coursey, Grace Hu, Sixian Li, Nadia Nahar, Shurui Zhou, Christian Kästner, and Jin L.C. Guo. 2023. Aspirations and Practice of ML Model Documentation: Moving the Needle with Nudging and Traceability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3581518

[34] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 401–413. https://doi.org/10.1145/3461702.3462571

[35] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 1119–1130. https://doi.org/10.18653/v1/D16-1120

[36] Su Lin Blodgett, Q. Vera Liao, Alexandra Olteanu, Rada Mihalcea, Michael Muller, Morgan Klaus Scheuerman, Chenhao Tan, and Qian Yang. 2022. Responsible Language Technologies: Foreseeing and Mitigating Harms. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3491101.3516502

[37] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. AI Design: A Responsible Artificial Intelligence Framework for Prefilling Impact Assessment Reports. *IEEE Internet Computing* 28, 5 (Sept. 2024), 37–45. https://doi.org/10.1109/MIC.2024.3451351

[38] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. Co-Designing an AI Impact Assessment Report Template with AI Practitioners and AI Compliance Experts. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 168–180. https://doi.org/10.1609/aies.v7i1.31627

[39] Edyta Bogucka, Sanja Šćepanović, and Daniele Quercia. 2024. Atlas of AI Risks: Enhancing Public Understanding of AI Risks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 12 (Oct. 2024), 33–43. https://doi.org/10.1609/hcomp.v12i1.31598

[40] Edyta Paulina Bogucka, Marios Constantinides, Julia De Miguel Velazquez, Sanja Scepanovic, Daniele Quercia, and Andrés Gvirtz. 2024. The Atlas of AI Incidents in Mobile Computing: Visualizing the Risks and Benefits of AI Gone Mobile. In *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction* (Melbourne, VIC, Australia) *(MobileHCI '24 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 26, 6 pages. https://doi.org/10.1145/3640471.3680447

[41] Mark C. Bolino and William H. Turnley. 1999. Measuring Impression Management in Organizations: A Scale Development Based on the Jones and Pittman Taxonomy. *Organizational Research Methods* 2, 2 (April 1999), 187–206. https://doi.org/10.1177/109442819922005

[42] Kenneth A. Bollen. 1989. *Structural Equations with Latent Variables*. John Wiley & Sons, New Jersey, United States.

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

[44] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. https://doi.org/10.48550/arXiv.2011.13416 arXiv:2011.13416 [cs]

[45] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[46] Virginia Braun and Victoria Clarke. 2012. Thematic Analysis. In *APA Handbook of Research Methods in Psychology, Vol 2: Research Designs: Quantitative, Qualitative, Neuropsychological, and Biological*. American Psychological Association, Washington, DC, US, 57–71. https://doi.org/10.1037/13620-004

[47] Virginia Braun and Victoria Clarke. 2021. *Thematic Analysis: A Practical Guide* (1st edition ed.). SAGE Publications Ltd, Los Angeles London New Delhi Singapore Washington DC Melbourne.

[48] J. Scott Brennen, Philip N. Howard, and Rasmus Kleis Nielsen. 2018. An Industry-Led Debate: How UK Media Cover Artificial Intelligence. https://doi.org/10.60625/RISJ-V219-D676

[49] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. https://doi.org/10.48550/arXiv.2306.03280 arXiv:2306.03280 [cs]

[50] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. Aha!: Facilitating ai impact assessment by generating examples of harms.

[51] Blake Bullwinkel, Amanda Minnich, Shiven Chawla, Gary Lopez, Martin Pouliot, Whitney Maxwell, Joris de Gruyter, Katherine Pratt, Saphir Qi, Nina Chikanov, et al. 2025. Lessons From Red Teaming 100 Generative AI Products.

[52] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.

[53] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[54] Zoraida Callejas, Brian Ravenet, Magalie Ochs, and Catherine Pelachaud. 2014. A Computational Model of Social Attitudes for a Virtual Recruiter. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 93–100.

[55] John P. Campbell. 1990. Modeling the Performance Prediction Problem in Industrial and Organizational Psychology. In *Handbook of Industrial and Organizational Psychology, Vol. 1, 2nd Ed.* Consulting Psychologists Press, Palo Alto, CA, US, 687–732.

[56] Michael A. Campion, David K. Palmer, and James E. Campion. 1997. A Review of Structure in the Selection Interview. *Personnel Psychology* 50, 3 (1997), 655–702. https://doi.org/10.1111/j.1744-6570.1997.tb00709.x

[57] Mark D. Cannon and Robert Witherspoon. 2005. Actionable Feedback: Unlocking the Power of Learning and Performance Improvement. *The Academy of Management Executive (1993-2005)* 19, 2 (2005), 120–134. arXiv:4166182

[58] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, Sunipa Dev, Vinodkumar Prabhakaran, David Ifeoluwa

Adelani, Dirk Hovy, and Luciana Benotti (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 53–67. https://doi.org/10.18653/v1/2023.c3nlp-1.7

[59] Rossella Cappetta and Chiara Paolino. 2015. Is It Always Worth Waiting? The Effect of Autonomy-supportive Teaching on Short-term and Long-term Learning Outcomes. *British Journal of Management* 26, 1 (2015), 93–108. https://doi.org/10.1111/1467-8551.12065

[60] Astrid Carolus, Martin J. Koch, Samantha Straka, Marc Erich Latoschik, and Carolin Wienrich. 2023. MAILS - Meta AI Literacy Scale: Development and Testing of an AI Literacy Questionnaire Based on Well-Founded Competency Models and Psychological Change- and Meta-Competencies. *Computers in Human Behavior: Artificial Humans* 1, 2 (Aug. 2023), 100014. https://doi.org/10.1016/j.chbah.2023.100014

[61] Christopher Carroll, Andrew Booth, Joanna Leaviss, and Jo Rick. 2013. "Best fit" framework synthesis: refining the method. *BMC medical research methodology* 13 (2013), 1–16.

[62] Jiyoo Chang and Christine Custis. 2022. Understanding Implementation Challenges in Machine Learning Documentation. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, Arlington VA USA, 1–8. https://doi.org/10.1145/3551624.3555301

[63] Maha Charfeddine, Habib M. Kammoun, Bechir Hamdaoui, and Mohsen Guizani. 2024. ChatGPT's Security Risks and Benefits: Offensive and Defensive Use-Cases, Mitigation Measures, and Future Implications. *IEEE Access* 12 (2024), 30263–30310. https://doi.org/10.1109/ACCESS.2024.3367792

[64] Sven Charleer, Francisco Gutiérrez, and Katrien Verbert. 2019. Supporting Job Mediator and Job Seeker through an Actionable Dashboard. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, Marina del Ray California, 121–131. https://doi.org/10.1145/3301275.3302312

[65] Beiwen Chen, Jasper Van Assche, Maarten Vansteenkiste, Bart Soenens, and Wim Beyers. 2015. Does Psychological Need Satisfaction Matter When Environmental or Financial Safety Are at Risk? *Journal of Happiness Studies* 16, 3 (June 2015), 745–766. https://doi.org/10.1007/s10902-014-9532-5

[66] Beiwen Chen, Maarten Vansteenkiste, Wim Beyers, Liesbet Boone, Edward L. Deci, Jolene Van der Kaap-Deeder, Bart Duriez, Willy Lens, Lennia Matos, Athanasios Mouratidis, Richard M. Ryan, Kennon M. Sheldon, Bart Soenens, Stijn Van Petegem, and Joke Verstuyf. 2015. Basic Psychological Need Satisfaction, Need Frustration, and Need Strength across Four Cultures. *Motivation and Emotion* 39, 2 (April 2015), 216–236. https://doi.org/10.1007/s11031-014-9450-1

[67] Canyu Chen and Kai Shu. 2024. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine* 45, 3 (2024), 354–368.

[68] Zhisheng Chen. 2023. Ethics and Discrimination in Artificial Intelligence-Enabled Recruitment Practices. *Humanities and Social Sciences Communications* 10, 1 (Sept. 2023), 1–12. https://doi.org/10.1057/s41599-023-02079-x

[69] Mauro Cherubini, Alex Jiahong Lu, Joey Chiao-Yin Hsiao, Muhan Zhao, Anandita Aggarwal, and Tawanna R Dillahunt. 2021. Elucidating Skills for Job Seekers: Insights and Critical Concerns from a Field Deployment in Switzerland. In *Designing Interactive Systems Conference 2021 (DIS '21)*. Association for Computing Machinery, New York, NY, USA, 449–465. https://doi.org/10.1145/3461778.3462049

[70] Marios Constantinides, Edyta Bogucka, Daniele Quercia, Susanna Kallio, and Mohammad Tahaei. 2024. RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (Nov. 2024), 388:1–388:28. https://doi.org/10.1145/3686927

[71] Marios Constantinides and Daniele Quercia. 2025. AI, Jobs, and the Automation Trap: Where Is HCI?. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work (CHIWORK '25)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3729176.3729191

[72] Kevin W. Cook, Carol A. Vance, and Paul E. Spector. 2000. The Relation of Candidate Personality With Selection-Interview Outcomes. *Journal of Applied Social Psychology* 30, 4 (2000), 867–885. https://doi.org/10.1111/j.1559-1816.2000.tb02828.x

[73] Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 25 July 2024. *Official Journal of the European Union* L 168 (2024), 1–15. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689 Accessed: 2024-07-29.

[74] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 427–439. https://doi.org/10.1145/3531146.3533108

[75] Tianyu Cui, Yanling Wang, Chuanpu Fu, Yong Xiao, Sijia Li, Xinhao Deng, Yunpeng Liu, Qinglin Zhang, Ziyi Qiu, Peiyang Li, Zhixing Tan, Junwu Xiong, Xinyu Kong, Zujie Wen, Ke Xu, and Qi Li. 2024. Risk Taxonomy, Mitigation, and Assessment Benchmarks of Large Language Model Systems. arXiv:2401.05778 [cs]

[76] Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and Unfairness in Information Retrieval Systems: New Challenges in the LLM Era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*. Association for Computing Machinery, New York, NY, USA, 6437–6447. https://doi.org/10.1145/3637528.3671458

[77] Ionut Damian, Tobias Baur, Birgit Lugrin, Patrick Gebhard, Gregor Mehlmann, and Elisabeth André. 2015. Games Are Better than Books: In-Situ Comparison of an Interactive Job Interview Game with Conventional Training. In *Artificial Intelligence in Education (Lecture Notes in Computer Science)*, Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa Verdejo (Eds.). Springer International Publishing, Cham, 84–94. https://doi.org/10.1007/978-3-319-19773-9_9

[78] Taufiq Daryanto, Xiaohan Ding, Lance T. Wilhelm, Sophia Stil, Kirk McInnis Knutsen, and Eugenia H. Rho. 2024. Conversate: Supporting Reflective Learning in Interview Practice Through Interactive Simulation and Dialogic Feedback. https://doi.org/10.48550/arXiv.2410.05570 arXiv:2410.05570 [cs]

[79] Dipto Das, Shion Guha, Jed R. Brubaker, and Bryan Semaan. 2024. The "Colonial Impulse" of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3613904.3642669

[80] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[81] Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American Language Bias in Natural Language Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6805–6824. https://doi.org/10.18653/v1/2023.emnlp-main.421

[82] Edward L. Deci and Richard M. Ryan. 1980. The Empirical Exploration of Intrinsic Motivational Processes. In *Advances in Experimental Social Psychology*, Leonard Berkowitz (Ed.). Vol. 13. Academic Press, Cambridge, Massachusetts, United States, 39–80. https://doi.org/10.1016/S0065-2601(08)60130-6

[83] Edward L. Deci and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4899-2271-7

[84] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. arXiv:2310.00907 [cs]

[85] Wesley Hanwen Deng, Solon Barocas, and Jennifer Wortman Vaughan. 2025. Supporting Industry Computing Researchers in Assessing, Articulating, and Addressing

the Potential Negative Societal Impact of Their Work. https://doi.org/10.1145/3711076 arXiv:2408.01057 [cs]

[86] Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, M. R. Leiser, and Saif Mohammad. 2023. Assessing Language Model Deployment with Risk Cards. arXiv:2303.18190 [cs]

[87] Leon Derczynski, Hannah Rose Kirk, Vidhisha Balachandran, Sachin Kumar, Yulia Tsvetkov, Mark R Leiser, and Saif Mohammad. 2023. Assessing language model deployment with risk cards.

[88] Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who Are the Stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Association for Computing Machinery, New York, NY, USA, 227–236. https://doi.org/10.1145/3514094.3534187

[89] Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2022. Theories of "Gender" in NLP Bias Research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Seoul Republic of Korea, 2083–2102. https://doi.org/10.1145/3531146.3534627

[90] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. 2023. Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation. *Information Fusion* 99 (Nov. 2023), 101896. https://doi.org/10.1016/j.inffus.2023.101896

[91] Tawanna Dillahunt and Chiao-Yin Hsiao. 2021. SkillsIdentifier: A Tool to Promote Career Identity and Self-efficacy Among Underrepresented Job Seekers. In *Proceedings of the Annual Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences, Honolulu, HI, USA, 4848–4857. https://doi.org/10.24251/HICSS.2021.588

[92] Tawanna R. Dillahunt, Nishan Bose, Suleman Diwan, and Asha Chen-Phang. 2016. Designing for Disadvantaged Job Seekers: Insights from Early Investigations. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16)*. Association for Computing Machinery, New York, NY, USA, 905–910. https://doi.org/10.1145/2901790.2901865

[93] Tawanna R. Dillahunt, Matthew Garvin, Marcy Held, and Julie Hui. 2021. Implications for Supporting Marginalized Job Seekers: Lessons from Employment Centers. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 324:1–324:24. https://doi.org/10.1145/3476065

[94] Tawanna R. Dillahunt and Joey Chiao-Yin Hsiao. 2020. Positive Feedback and Self-Reflection: Features to Support Self-efficacy among Underrepresented Job Seekers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*

(CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376717

[95] Tawanna R. Dillahunt, Jason Lam, Alex Lu, and Earnest Wheeler. 2018. Designing Future Employment Applications for Underserved Job Seekers: A Speed Dating Study. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/3196709.3196770

[96] Tawanna R Dillahunt, Lucas Siqueira Rodrigues, Joey Chiao-Yin Hsiao, and Mauro Cherubini. 2022. Self-Regulation and Autonomy in the Job Search: Key Factors to Support Job Search Among Swiss Job Seekers. *Interacting with Computers* 33, 5 (May 2022), iwac008. https://doi.org/10.1093/iwc/iwac008

[97] Pierre Dillenbourg, Sanna Järvelä, and Frank Fischer. 2009. The Evolution of Research on Computer-Supported Collaborative Learning. In *Technology-Enhanced Learning: Principles and Products*, Nicolas Balacheff, Sten Ludvigsen, Ton de Jong, Ard Lazonder, and Sally Barnes (Eds.). Springer Netherlands, Dordrecht, 3–19. https://doi.org/10.1007/978-1-4020-9827-7_1

[98] Kimberly Do, Rock Yuren Pang, Jiachen Jiang, and Katharina Reinecke. 2023. "That's Important, but...": How Computer Science Researchers Anticipate Unintended Consequences of Their Research Innovations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581347

[99] Upol Ehsan, Q. Vera Liao, Samir Passi, Mark O. Riedl, and Hal Daumé. 2024. Seamful XAI: Operationalizing Seamful Design in Explainable AI. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 119:1–119:29. https://doi.org/10.1145/3637396

[100] Sheila Embleton. 2023. Names in India: History, Colonialism, Renaming, Contemporary Issues. In *Proceedings of the 27th International Congress of Onomastic Sciences*. Jagiellonian University Press. https://doi.org/10.4467/K7501.45/22.23.18048

[101] OpenAI et al. 2024. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs]

[102] Social Affairs European Labour Authority, Directorate-General for Employment and Inclusion. 2021. How COVID-19 Has Fundamentally Changed Recruitment. https://eures.ec.europa.eu/how-covid-19-has-fundamentally-changed-recruitment-2021-10-13_en Online; accessed 10 July 2023.

[103] Sabri Eyuboglu, Karan Goel, Arjun Desai, Lingjiao Chen, Mathew Monfort, Chris Ré, and James Zou. 2024. Model ChangeLists: Characterizing Updates to ML Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 2432–2453. https://doi.org/10.1145/3630106.3659047

[104] Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J. Biega. 2025. Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey. *ACM Trans. Intell. Syst. Technol.* 16, 1 (Jan. 2025), 16:1–16:54. https://doi.org/10.1145/3696457

[105] Alessandro Fabris, Nina Baranowska, Matthew J. Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik J. Zuiderveen Borgesius, and Asia J. Biega. 2025. Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey. *ACM Transactions on Intelligent Systems and Technology* 16, 1 (2025), 16:1–16:54.

[106] Morten W Fagerland, Stian Lydersen, and Petter Laake. 2015. Recommended Confidence Intervals for Two Independent Binomial Proportions. *Statistical Methods in Medical Research* 24, 2 (April 2015), 224–254. https://doi.org/10.1177/0962280211415469

[107] Huanming Fang, Hui Miao, Karan Shukla, Dan Nanas, Catherina Xu, Christina Greer, Neoklis Polyzotis, Tulsee Doshi, Tiffany Deng, Margaret Mitchell, et al. 2020. Introducing the model card toolkit for easier model transparency reporting.

[108] Jiing-Lih Farh and W. E. Scott. 1983. The Experimental Effects of "Autonomy" on Performance and Self-Reports of Satisfaction. *Organizational Behavior and Human Performance* 31, 2 (April 1983), 203–222. https://doi.org/10.1016/0030-5073(83)90121-6

[109] Enrique Fernández-Macías. 2018. Automation, digitisation and platforms: implications for work and employment. https://www.eurofound.europa.eu/publications/report/2018/automation-digitisation-and-platforms-implications-for-work-and-employment

[110] Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic Bias in ChatGPT: Language Models Reinforce Dialect Discrimination. arXiv:2406.08818 [cs.CL] https://arxiv.org/abs/2406.08818

[111] Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* 32, 3 (1948), 221–233.

[112] Jodi Forlizzi. 2018. Moving beyond User-Centered Design. *interactions* 25, 5 (Aug. 2018), 22–23. https://doi.org/10.1145/3239558

[113] Lynn Frewer. 2004. The Public and Effective Risk Communication. *Toxicology Letters* 149, 1-3 (April 2004), 391–397. https://doi.org/10.1016/j.toxlet.2003.12.049

[114] Joseph B Fuller, Manjari Raman, Eva Sage-Gavin, and Kristen Hines. 2021. Hidden Workers: Untapped Talent. *Harvard Business School* (2021).

[115] Michelle Fung, Yina Jin, RuJie Zhao, and Mohammed (Ehsan) Hoque. 2015. ROC Speak: Semi-Automated Personalized Feedback on Nonverbal Behavior from Recorded

Videos. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, Osaka Japan, 1167–1178. https://doi.org/10.1145/2750858.2804265

[116] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997 [cs]

[117] Kiran Garimella and Dean Eckles. 2017. Image based misinformation on WhatsApp. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)*.

[118] Patrick Gebhard, Tobias Baur, Ionut Damian, Gregor Mehlmann, Johannes Wagner, and Elisabeth André. 2014. Exploring Interaction Strategies for Virtual Characters to Induce Stress in Simulated Job Interviews. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 661–668.

[119] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[120] Team Gemini. 2025. Gemini: A Family of Highly Capable Multimodal Models. https://doi.org/10.48550/arXiv.2312.11805 arXiv:2312.11805 [cs]

[121] Theodoros Georgiou, Lynne Baillie, Martin K. Ross, and Frank Broz. 2020. Applying the Participatory Design Workshop Method to Explore How Socially Assistive Robots Could Assist Stroke Survivors. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Cambridge United Kingdom, 203–205. https://doi.org/10.1145/3371382.3378232

[122] Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2024. Do Generative AI Models Output Harm While Representing Non-Western Cultures: Evidence from A Community-Centered Approach. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (Oct. 2024), 476–489. https://doi.org/10.1609/aies.v7i1.31651

[123] Debra K. Godfrey, Edward E. Jones, and Charles G. Lord. 1986. Self-Promotion Is Not Ingratiating. *Journal of Personality and Social Psychology* 50, 1 (1986), 106–115. https://doi.org/10.1037/0022-3514.50.1.106

[124] Abenezer Golda, Kidus Mekonen, Amit Pandey, Anushka Singh, Vikas Hassija, Vinay Chamola, and Biplab Sikdar. 2024. Privacy and Security Concerns in Generative AI: A Comprehensive Survey. *IEEE Access* 12 (2024), 53485–53499. https://doi.org/10.1109/ACCESS.2024.3381611

[125] Delaram Golpayegani, Isabelle Hupont, Cecilia Panigutti, Harshvardhan J. Pandit, Sven Schade, Declan O'Sullivan, and Dave Lewis. 2024. AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act. In *Privacy Technologies and Policy: 12th Annual Privacy Forum, APF 2024, Karlstad, Sweden, September 4–5, 2024, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 48–72. https://doi.org/10.1007/978-3-031-68024-3_3

[126] Delaram Golpayegani, Harshvardhan J. Pandit, and Dave Lewis. 2023. To Be High-Risk, or Not To Be—Semantic Specifications and Implications of the AI Act's High-Risk AI Applications and Harmonised Standards. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 905–915. https://doi.org/10.1145/3593013.3594050

[127] Google. 2024. Model cards: A step towards AI transparency. https://modelcards.withgoogle.com/about. Accessed: 2024-07-29.

[128] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field Methods* 18, 1 (2006), 59–82. https://doi.org/10.1177/1525822X05279903

[129] Michael Hameleers, Thomas E Powell, Toni GLA Van Der Meer, and Lieke Bos. 2020. A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political communication* 37, 2 (2020), 281–301.

[130] Dominik Hangartner, Daniel Kopp, and Michael Siegenthaler. 2021. Monitoring Hiring Discrimination through Online Recruitment Platforms. *Nature* 589, 7843 (Jan. 2021), 572–576. https://doi.org/10.1038/s41586-020-03136-0

[131] Christopher J. Hartwell, Clark D. Johnson, and Richard A. Posthuma. 2019. Are We Asking the Right Questions? Predictive Validity Comparison of Four Structured Interview Question Types. *Journal of Business Research* 100, September 2018 (2019), 122–129. https://doi.org/10.1016/j.jbusres.2019.03.026

[132] Felix Hausdorff. 2005. *Set Theory*. American Mathematical Soc.

[133] Andrew F. Hayes. 2013. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. Guilford Press, New York, NY, US. xvii, 507 pages.

[134] Changyang He, Yue Deng, Alessandro Fabris, Bo Li, and Asia Biega. 2025. Developing a Fair Online Recruitment Framework Based on Job-seekers' Fairness Concerns. https://doi.org/10.48550/arXiv.2501.14110 arXiv:2501.14110 [cs]

[135] Léo Hemamou, Ghazi Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. 2019. HireNet: A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews. In *Proceedings of the AAAI Conference on*

Artificial Intelligence, Vol. 33. AAAI Publications, Washington, DC, USA, 573–581. https://doi.org/10.1609/aaai.v33i01.3301573

[136] David G. Hendry, Norah Abokhodair, Rose Paquet Kinsley, and Jill Palzkill Woelfer. 2017. Homeless Young People, Jobs, and a Future Vision: Community Members' Perceptions of the Job Co-op. In *Proceedings of the 8th International Conference on Communities and Technologies (C&T '17)*. Association for Computing Machinery, New York, NY, USA, 22–31. https://doi.org/10.1145/3083671.3083680

[137] Viviane Herdel, Sanja Šćepanović, Edyta Bogucka, and Daniele Quercia. 2024. ExploreGen: Large Language Models for Envisioning the Uses and Risks of AI Technologies. https://doi.org/10.48550/arXiv.2407.12454 arXiv:2407.12454 [cs]

[138] Catherine M. Hicks, Vineet Pandey, C. Ailie Fraser, and Scott Klemmer. 2016. Framing Feedback: Choosing Review Environment Features That Support High Quality Peer Assessment. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 458–469. https://doi.org/10.1145/2858036.2858195

[139] Geert Hofstede. 1980. Culture and Organizations. *International Studies of Management & Organization* 10, 1 (Dec. 1980), 15–41.

[140] Geert Hofstede. 2011. Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture* 2, 1 (Dec. 2011), 1–26. https://doi.org/10.9707/2307-0919.1014

[141] Jess Hohenstein, Rene F. Kizilcec, Dominic DiFranzo, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeffrey Hancock, and Malte F. Jung. 2023. Artificial Intelligence in Communication Impacts Language and Social Relationships. *Scientific Reports* 13, 1 (April 2023), 5487. https://doi.org/10.1038/s41598-023-30938-9

[142] James G. Hollandsworth, Mary E. Dressel, and Joanne Stevens. 1977. Use of behavioral versus traditional procedures for increasing job interview skills. *Journal of Counseling Psychology* 24, 6 (1977), 503–510. https://doi.org/10.1037/0022-0167.24.6.503

[143] Mohammed (Ehsan) Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. 2013. MACH: My Automated Conversation Coach. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 697–706. https://doi.org/10.1145/2493432.2493502

[144] Ryan Horn and Tara Behrend. 2017. Video Killed the Interview Star: Does Picture-in-Picture Affect Interview Performance? *Personnel Assessment and Decisions* 3, 1 (2017), 51–59. https://doi.org/10.25035/pad.2017.005

[145] The White House. 2022. Blueprint for an AI Bill of Rights | OSTP. https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[146] Allen Huffcutt, Chad Van Iddekinge, and Philip Roth. 2011. Understanding applicant behavior in employment interviews: A theoretical model of interviewee performance. *Human Resource Management Review - HUMAN RESOURCE MANAGEMENT REV* 21 (06 2011), 353–367. https://doi.org/10.1016/j.hrmr.2011.05.003

[147] Allen I Huffcutt, James M Conway, Nancy J Stone, and Philip L Roth. 2001. Identification and Meta-Analytic Assessment of Psychological Constructs Measured in Employment Interviews. *Journal of Applied Psychology* 86, 5 (2001), 897–913. https://doi.org/10.1037//0021-9010.86.5.897

[148] Isabelle Hupont, David Fernández-Llorca, Sandra Baldassarri, and Emilia Gómez. 2024. Use Case Cards: A Use Case Reporting Framework Inspired by the European AI Act. *Ethics and Information Technology* 26, 2 (March 2024), 19. https://doi.org/10.1007/s10676-024-09757-7

[149] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 560–575. https://doi.org/10.1145/3442188.3445918

[150] Ken Hyland. 2011. Metadiscourse: Mapping interactions in academic writing. *Journal of Pragmatics* 43, 11 (2011), 1757–1770.

[151] Andrew S. Imada and Milton D. Hakel. 1978. Influence of Nonverbal Communication and Rater Proximity on Impressions and Decisions in Simulated Employment Interviews. *Journal of Applied Psychology* 62, 3 (Jan. 1978), 295. https://doi.org/10.1037/0021-9010.62.3.295

[152] International Labour Organization. 2024. World Employment and Social Outlook: May 2024 Update | International Labour Organization. https://www.ilo.org/publications/flagship-reports/world-employment-and-social-outlook-may-2024-update.

[153] M Jaakkola. 2023. A handbook for journalism educators: Reporting on Artificial Intelligence. *United Nations Educational, Scientific and Cultural Organization* (2023). https://unesdoc.unesco.org/ark:/48223/pf0000384551

[154] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 624–635. https://doi.org/10.1145/3442188.3445923

[155] Rebecca L. Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The Ghost in the Machine Has an American Accent: Value Conflict in GPT-3. https://doi.org/10.48550/arXiv.2203.07785 arXiv:2203.07785 [cs]

[156] Ruth Ellen Jones and Kareem R. Abdelfattah. 2020 Jul - Aug. Virtual Interviews in the Era of COVID-19: A Primer for Applicants. *Journal of Surgical Education* 77, 4 (2020 Jul - Aug), 733–734. https://doi.org/10.1016/j.jsurg.2020.03.020

[157] Aparna Joshi, David A. Bloom, Amy Spencer, Kara Gaetke-Udager, and Richard H. Cohan. 2020. Video Interviewing: A Review and Recommendations for Implementation in the Era of COVID-19 and Beyond. *Academic Radiology* 27, 9 (Sept. 2020), 1316–1322. https://doi.org/10.1016/j.acra.2020.05.020

[158] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large Language Models Struggle to Learn Long-Tail Knowledge. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Hawai'i, 15696–15707.

[159] Jasmine Kaur, Rishabh Devgon, and Dhairya Chaudhary. 2022. Work of Fiction: Using Speculative Design to Deliberate on the Future of Hiring. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing (CSCW'22 Companion)*. Association for Computing Machinery, New York, NY, USA, 23–26. https://doi.org/10.1145/3500868.3559442

[160] Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. 2024. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder Early-stage Deliberations Around Public Sector AI Proposals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–22. https://doi.org/10.1145/3613904.3642849 arXiv:2402.18774 [cs]

[161] DeBrae Kennedy-Mayo and Jake Gord. 2024. "Model Cards for Model Reporting" in 2024: Reclassifying Category of Ethical Considerations in Terms of Trustworthiness and Risk Management. https://doi.org/10.48550/arXiv.2403.15394 arXiv:2403.15394 [cs]

[162] Khyati Khandelwal, Manuel Tonneau, Andrew M. Bean, Hannah Rose Kirk, and Scott A. Hale. 2024. Indian-BhED: A Dataset for Measuring India-Centric Biases in Large Language Models. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*. ACM, Bremen Germany, 231–239. https://doi.org/10.1145/3677525.3678666

[163] Jieun Kim, Hauke Sandhaus, and Susan R. Fussell. 2023. VR Job Interview Using a Gender-Swapped Avatar. In *Computer Supported Cooperative Work and Social Computing*. ACM, Minneapolis MN USA, 154–159. https://doi.org/10.1145/3584931.3606976

[164] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Chicago IL USA, 77–88. https://doi.org/10.1145/3593013.3593978

[165] Taewan Kim, Mintra Ruensuk, and Hwajung Hong. 2020. In Helping a Vulnerable Bot, You Help Yourself: Designing a Social Bot as a Care-Receiver to Promote Mental Health and Reduce Stigma. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376743

[166] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[167] Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine* 15, 2 (June 2016), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

[168] Emily Jane Kothe and Mathew Ling. 2019. *Retention of participants recruited to a multi-year longitudinal study via Prolific*. preprint. PsyArXiv. https://doi.org/10.31234/osf.io/5yv2u

[169] Deepak Kumar. 2020. *Language Discrimination in Indian Higher Education*. Lokmitra Publication, 148–169.

[170] Deepak Kumar, Tessa Grosz, Navid Rekabsaz, Elisabeth Greif, and Markus Schedl. 2023. Fairness of Recommender Systems in the Recruitment Domain: An Analysis from Technical and Legal Perspectives. *Frontiers in Big Data* 6 (Oct. 2023). https://doi.org/10.3389/fdata.2023.1245198

[171] Sachin Kumar, Vidhisha Balachandran, Lucille Njoo, Antonios Anastasopoulos, and Yulia Tsvetkov. 2023. Language Generation Models Can Cause Harm: So What Can We Do About It? An Actionable Survey. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 3299–3321. https://doi.org/10.18653/v1/2023.eacl-main.241

[172] Markus Langer, Cornelius J. König, and Andromachi Fitili. 2018. Information as a Double-Edged Sword: The Role of Computer Experience and Information on Applicant Reactions towards Novel Technologies for Personnel Selection. *Computers in Human Behavior* 81 (April 2018), 19–30. https://doi.org/10.1016/j.chb.2017.11.036

[173] Markus Langer, Cornelius J. König, Patrick Gebhard, and Elisabeth André. 2016. Dear Computer, Teach Me Manners: Testing Virtual Employment Interview Training. *International Journal of Selection and Assessment* 24, 4 (2016), 312–323. https://doi.org/10.1111/ijsa.12150

[174] Markus Langer, Cornelius J König, and Kevin Krause. 2017. Examining Digital Interviews for Personnel Selection: Applicant Reactions and Interviewer Ratings. *International Journal of Selection and Assessment* 25, 4 (2017), 371–382. https://doi.org/10.1111/ijsa.12191

[175] Mitra Lashkari and Jinghui Cheng. 2023. "Finding the Magic Sauce": Exploring Perspectives of Recruiters and Job Seekers on Recruitment Bias and Automated Tools. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. https://doi.org/10.1145/3544548.3581548

[176] Van M. Latham. 1987. Interviewee Training: A Review of Some Empirical Literature. *Journal of Career Development* 14, 2 (1987), 96–107. https://doi.org/10.1177/089484538701400204 arXiv:https://doi.org/10.1177/089484538701400204

[177] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12, 3 (2022), e1452.

[178] Hee Rin Lee, Selma Šabanović, Wan-Ling Chang, Shinichi Nagata, Jennifer Piatt, Casey Bennett, and David Hakken. 2017. Steps Toward Participatory Design of Social Robots: Mutual Learning with Older Adults with Depression. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Vienna Austria, 244–253. https://doi.org/10.1145/2909824.3020237

[179] Minha Lee, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for Vincent: A Chatbot for Self-Compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300932

[180] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–35. https://doi.org/10.1145/3359283

[181] Yang W Lee, Diane M Strong, Beverly K Kahn, and Richard Y Wang. 2002. AIMQ: a methodology for information quality assessment. *Information & management* 40, 2 (2002), 133–146.

[182] Mark R. Lepper, David Greene, and Richard E. Nisbett. 1973. Undermining Children's Intrinsic Interest with Extrinsic Reward: A Test of the "Overjustification" Hypothesis. *Journal of Personality and Social Psychology* 28 (1973), 129–137. https://doi.org/10.1037/h0035519

[183] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., Vancouver, Canada, 9459–9474.

[184] Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. 2024. Systematic Analysis of 32,111 AI Model Cards Characterizes Documentation Practice in AI. *Nature Machine Intelligence* 6, 7 (July 2024), 744–753. https://doi.org/10.1038/s42256-024-00857-z

[185] Q. Vera Liao, Hariharan Subramonyam, Jennifer Wang, and Jennifer Wortman Vaughan. 2023. Designerly Understanding: Information Needs for Model Transparency to Support Design Ideation for AI-Powered User Experience. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3580652

[186] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81.

[187] Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024. Preventing and detecting misinformation generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3001–3004.

[188] Jiarui Liu, Wenkai Li, Zhijing Jin, and Mona Diab. 2024. Automatic Generation of Model and Data Cards: A Step Towards Responsible AI. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1975–1997.

[189] Jiarui Liu, Wenkai Li, Zhijing Jin, and Mona Diab. 2024. Automatic Generation of Model and Data Cards: A Step Towards Responsible AI. arXiv:2405.06258 [cs]

[190] Yihao Liu, Jinhe Huang, Yanjie Li, Dong Wang, and Bin Xiao. 2025. Generative AI Model Privacy: A Survey. *Artificial Intelligence Review* 58, 33 (2025), 1–25. https://doi.org/10.1007/s10462-024-11024-6

[191] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2024. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374 [cs.AI] https://arxiv.org/abs/2308.05374

[192] LMSYS. 2024. *LMSYS Chatbot Arena Leaderboard*. Huggingface. Retrieved July 9, 2024 from https://lmsys.org/blog/2023-06-22-leaderboard/

[193] Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's Only a Computer: Virtual Humans Increase Willingness to Disclose. *Computers in Human Behavior* 37 (Aug. 2014), 94–100. https://doi.org/10.1016/j.chb.2014.04.043

[194] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems* 36 (2024).

[195] Eden-Raye Lukacik, Joshua S. Bourdage, and Nicolas Roulin. 2020. Into the Void: A Conceptual Model and Research Agenda for the Design and Use of Asynchronous Video Interviews. *Human Resource Management Review* 32, 1 (Oct. 2020), 100789. https://doi.org/10.1016/j.hrmr.2020.100789

[196] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[197] Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. 2023. Intersectional Stereotypes in Large Language Models: Dataset and Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 8589–8597. https://doi.org/10.18653/v1/2023.findings-emnlp.575

[198] Jonne Maas and Aarón Moreno Inglés. 2024. Beyond Participatory AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 932–942. https://doi.org/10.1609/aies.v7i1.31693

[199] Bill MacCartney and Christopher D. Manning. 2008. Modeling Semantic Containment and Exclusion in Natural Language Inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Donia Scott and Hans Uszkoreit (Eds.). Coling 2008 Organizing Committee, Manchester, UK, 521–528.

[200] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1 (April 2022), 52:1–52:26. https://doi.org/10.1145/3512899

[201] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[202] Juan M. Madera, Michelle R. Hebl, and Randi C. Martin. 2009. Gender and Letters of Recommendation for Academia: Agentic and Communal Differences. *Journal of Applied Psychology* 94, 6 (2009), 1591–1599. https://doi.org/10.1037/a0016539

[203] Alessandro Mantelero and Maria Samantha Esposito. 2021. An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law & Security Review* 41 (2021), 105561.

[204] Frank Martela and Richard M. Ryan. 2016. Prosocial Behavior Increases Well-Being and Vitality Even without Contact with the Beneficiary: Causal and Behavioral Evidence. *Motivation and Emotion* 40, 3 (June 2016), 351–357. https://doi.org/10.1007/s11031-016-9552-z

[205] Todd J Maurer, Jerry M Solamon, Kimberly D Andrews, and Deborah D Troxtel. 2001. Interviewee coaching, preparation strategies, and response strategies in relation to performance in situational employment interviews: An extension of Maurer, Solamon, and Troxtel (1998). *Journal of Applied Psychology* 86, 4 (2001), 709.

[206] Julie McCarthy and Richard Goffin. 2004. Measuring Job Interview Anxiety: Beyond Weak Knees and Sweaty Palms. *Personnel Psychology* 57, 3 (2004), 607–637. https://doi.org/10.1111/j.1744-6570.2004.00002.x

[207] Sean McGregor. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 15458–15463. https://doi.org/10.1609/aaai.v35i17.17817

[208] Scott McLean, Gemma JM Read, Jason Thompson, Chris Baber, Neville A Stanton, and Paul M Salmon. 2023. The risks associated with Artificial General Intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence* 35, 5 (2023), 649–663.

[209] Thomas Mejtoft, Sarah Hale, and Ulrik Söderström. 2019. Design Friction. In *Proceedings of the 31st European Conference on Cognitive Ergonomics (ECCE '19)*. Association for Computing Machinery, New York, NY, USA, 41–44. https://doi.org/10.1145/3335082.3335106

[210] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented Language Models: A Survey. arXiv:2302.07842 [cs]

[211] Matthew B. Miles, A. Michael Huberman, and Johnny Saldana. 2013. *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications.

[212] Matthew B. Miles, A. Michael Huberman, and Johnny Saldana. 2013. *Qualitative Data Analysis: A Methods Sourcebook*. SAGE Publications, CA, USA.

[213] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model

Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 220–229. https://doi.org/10.1145/3287560.3287596

[214] Brent Mittelstadt, Sandra Wachter, and Chris Russell. 2023. To protect science, we must use LLMs as zero-shot translators. *Nature Human Behaviour* 7, 11 (2023), 1830–1832.

[215] Saif M Mohammad. 2022. Ethics sheet for AI tasks. *Computational Linguistics* 48, 2 (2022), 239–278.

[216] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2014–2037. https://doi.org/10.18653/v1/2023.eacl-main.148

[217] Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, Local Biases: Socio-Cultural Prejudices across Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 15828–15845. https://doi.org/10.18653/v1/2023.emnlp-main.981

[218] Skanda Muralidhar, Laurent Son Nguyen, Denise Frauendorfer, Jean-Marc Odobez, Marianne Schmid Mast, and Daniel Gatica-Perez. 2016. Training on the Job: Behavioral Analysis of Job Interviews in Hospitality. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 84–91. https://doi.org/10.1145/2993148.2993191

[219] Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. 2015. Automated Prediction and Analysis of Job Interview Performance: The Role of What You Say and How You Say It. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Vol. 1. IEEE, New York, NY, USA, 1–6. https://doi.org/10.1109/FG.2015.7163127

[220] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Towards Involving End-users in Interactive Human-in-the-loop AI Fairness. https://doi.org/10.48550/arXiv.2204.10464 arXiv:2204.10464

[221] Tarek Naous and Wei Xu. 2025. On the Origin of Cultural Biases in Language Models: From Pre-training Data to Linguistic Phenomena. In *Proceedings of the 2025 Conference of the North American Chapter of the ACL: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics, 6423–6443.

[222] Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality Bias in Text Generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 116–122. https://doi.org/10.18653/v1/2023.eacl-main.9

[223] National Institute of Standards and Technology (NIST). 2023. AI Risk Management Framework. https://www.nist.gov/itl/ai-risk-management-framework Accessed: 2024-11-19.

[224] Laurent Son Nguyen, Denise Frauendorfer, Marianne Schmid Mast, and Daniel Gatica-Perez. 2014. Hire Me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior. *IEEE Transactions on Multimedia* 16, 4 (June 2014), 1018–1031. https://doi.org/10.1109/TMM.2014.2307169

[225] Laurent Son Nguyen and Daniel Gatica-Perez. 2016. Hirability in the Wild: Analysis of Online Conversational Video Resumes. *IEEE Transactions on Multimedia* 18, 7 (July 2016), 1422–1437. https://doi.org/10.1109/TMM.2016.2557058

[226] NIST. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0).

[227] NIST. 2024. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.

[228] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. 2024. Accountability in Artificial Intelligence: What It Is and How It Works. *AI & SOCIETY* 39, 4 (Aug. 2024), 1871–1882. https://doi.org/10.1007/s00146-023-01635-y

[229] José Luiz Nunes, Gabriel D. J. Barbosa, Clarisse Sieckenius de Souza, Helio Lopes, and Simone D. J. Barbosa. 2022. Using Model Cards for Ethical Reflection: A Qualitative Exploration. In *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems (IHC '22)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3554364.3559117

[230] OECD. [n. d.]. Participatory AI Framework. https://oecd.ai/en/catalogue/tools/participatory-ai-framework.

[231] OECD. 2019. *OECD Employment Outlook 2019: The Future of Work*. Organisation for Economic Co-operation and Development, Paris France.

[232] OECD. 2020. Guidelines for MNEs - Organisation for Economic Co-operation and Development. https://mneguidelines.oecd.org/rbc-and-digitalisation.htm.

[233] OECD. 2021. *OECD Employment Outlook 2021: Navigating the COVID-19 Crisis and Recovery*. Organisation for Economic Co-operation and Development, Paris France.

[234] OECD. 2022. *OECD Employment Outlook 2022*. Organisation for Economic Co-operation and Development, Paris France.

[235] OECD. 2023. *OECD Employment Outlook 2023: Artificial Intelligence and the Labour Market*. Organisation for Economic Co-operation and Development, Paris.

[236] OECD. 2024. AI Incidents. https://oecd.ai/en/incidents Accessed: 2024-11-27.

[237] Michio Okada. 2023. Weak Robots: Relational-Oriented Approach to Human Well-being. In *Systems Design Based on the Benefits of Inconvenience*, Hiroshi Kawakami (Ed.). Springer Nature, Singapore, 57–68. https://doi.org/10.1007/978-981-19-9588-0_6

[238] Organisation for Economic Co-operation and Development (OECD). 2025. OECD AI Incidents Database. https://oecd.ai/en/incidents-methodology online at https://oecd.ai/en/incidents-methodology. Accessed: August 2025.

[239] Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing Toxic Content in Large Pre-Trained Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4262–4274. https://doi.org/10.18653/v1/2021.acl-long.329

[240] Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. https://doi.org/10.48550/arXiv.2312.05934 arXiv:2312.05934 [cs]

[241] Sinead O'Connor and Helen Liu. 2024. Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & SOCIETY* 39, 4 (2024), 2045–2057.

[242] Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models?

[243] Monica Pereira and Kate Hone. 2021. Communication Skills Training Intervention Based on Automated Recognition of Nonverbal Signals. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445324

[244] Karine Perset and Luis Aranda. 2024. *Defining AI Incidents and Related Terms*. OECD Artificial Intelligence Papers No. 16. OECD. https://www.oecd.ai/incidents Approved by the OECD Digital Policy Committee on 14 March 2024.

[245] David Piorkowski, Michael Hind, and John Richards. 2024. Quantitative AI Risk Assessments: Opportunities and Challenges. https://doi.org/10.48550/arXiv.2209.06317 arXiv:2209.06317 [cs]

[246] Adriana Placani. 2024. Anthropomorphism in AI: Hype and Fallacy. *AI and Ethics* (Feb. 2024). https://doi.org/10.1007/s43681-024-00419-4

[247] Ritika Poddar, Rashmi Sinha, Mor Naaman, and Maurice Jakesch. 2023. AI Writing Assistants Influence Topic Choice in Self-Presentation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3544549.3585893

[248] Charlie Pownall. 2023. AI, Algorithmic and Automation Incident and Controversy Repository (AIAAIC). https://www.aiaaic.org/ online at https://www.aiaaic.org/. Accessed: August 2025.

[249] T. Alexander Puutio and Patrick K. Lin. 2025. First Come, First Hired? ChatGPT's Bias for The First Resume It Sees and the Cost for Candidates to Overcome Bias in AI Hiring Tools. arXiv:5146809

[250] Rida Qadri, Renee Shelby, Cynthia L. Bennett, and Emily Denton. 2023. AI's Regimes of Representation: A Community-centered Study of Text-to-Image Models in South Asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 506–517. https://doi.org/10.1145/3593013.3594016

[251] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/arXiv.2212.04356 arXiv:2212.04356 [eess]

[252] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 469–481. https://doi.org/10.1145/3351095.3372828

[253] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

[254] Pooja S. B. Rao, Manish Agnihotri, and Dinesh Babu Jayagopi. 2021. Improving Asynchronous Interview Interaction with Follow-up Question Generation. *International Journal of Interactive Multimedia and Artificial Intelligence* 6, Special Issue on Artificial Intelligence, Paving the Way to the Future (2021), 79–89.

[255] Pooja S. B. Rao, Dinesh Babu Jayagopi, and Mauro Cherubini. 2022. Explainability in Automated Training and Feedback Systems. https://doi.org/10.31219/osf.io/qd87c

[256] Pooja S. B. Rao, Sowmya Rasipuram, Rahul Das, and Dinesh Babu Jayagopi. 2017. Automatic Assessment of Communication Skill in Non-Conventional Interview Settings: A Comparative Study. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI '17)*. Association for Computing Machinery, New York, NY, USA, 221–229. https://doi.org/10.1145/3136755.3136756

[257] Pooja S. B. Rao, Laetitia Renier, Marc-Olivier Boldi, Marianne Schmid Mast, Dinesh Babu Jayagopi, and Mauro Cherubini. 2024. On the Potential of Supporting Autonomy in Online Video Interview Training Platforms. *International Journal of Human-Computer Studies* 191 (Nov. 2024), 103326. https://doi.org/10.1016/j.ijhcs.2024.103326

[258] Pooja S. B. Rao, Sanja Šćepanović, Ke Zhou, Edyta Paulina Bogucka, and Daniele Quercia. 2025. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–26. https://doi.org/10.1145/3706598.3713979

[259] Pooja S. B. Rao, Sanja Šćepanović, Ke Zhou, Edyta Paulina Bogucka, and Daniele Quercia. 2025. RiskRAG: A Data-Driven Solution for Improved AI Model Risk Reporting. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 11, 26 pages. https://doi.org/10.1145/3706598.3713979

[260] Sowmya Rasipuram and Dinesh Babu Jayagopi. 2020. Automatic Multimodal Assessment of Soft Skills in Social Interactions: A Review. *Multimedia Tools and Applications* 79, 19 (May 2020), 13037–13060. https://doi.org/10.1007/s11042-019-08561-6

[261] Sowmya Rasipuram, Pooja Rao, and Dinesh Babu Jayagopi. 2016. Asynchronous Video Interviews vs. Face-To-face Interviews for Communication Skill Measurement: A Systematic Study. In *ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 370–377. https://doi.org/10.1145/2993148.2993183

[262] Charvi Rastogi, Marco Tulio Ribeiro, Nicholas King, Harsha Nori, and Saleema Amershi. 2023. Supporting Human-AI Collaboration in Auditing LLMs with LLMs. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 913–926. https://doi.org/10.1145/3600211.3604712

[263] Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Ramona Comanescu, Canfer Akbulut, Tom Stepleton, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, William Isaac, and Laura Weidinger. 2024. Gaps in the Safety Evaluation of Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, 1 (Oct. 2024), 1200–1217. https://doi.org/10.1609/aies.v7i1.31717

[264] Athena Richards, Sheila Convery, Margaret O'Mahony, and Brian Caulfield. 2024. Pre and post Covid preferences for working from home. *Travel Behaviour and Society* 34 (2024), 100679.

[265] C. Scott Rigby and Andrew K. Przybylski. 2009. Virtual Worlds and the Learner Hero: How Today's Video Games Can Inform Tomorrow's Digital Learning Environments.

Theory and Research in Education 7, 2 (July 2009), 214–223. https://doi.org/10.1177/1477878509104326

[266] Scott Rigby and Richard M. Ryan. 2011. *Glued to Games: How Video Games Draw Us In and Hold Us Spellbound: How Video Games Draw Us In and Hold Us Spellbound.* ABC-CLIO, Santa Barbara, California.

[267] Kortney A Robinson, Borami Shin, and Sidhu P Gangadharan. 2021. A Comparison Between In-Person and Virtual Fellowship Interviews During the COVID-19 Pandemic. *Journal of Surgical Education* 78, 4 (July 2021), 1175–1181. https://doi.org/10.1016/j.jsurg.2020.11.006

[268] Rowena Rodrigues, Anais Resseguier, and Nicole Santiago. 2023. When Artificial Intelligence Fails: The Emerging Role of Incident Databases. *Pub. Governance, Admin. & Fin. L. Rev.* 8 (2023), 17.

[269] Victoria Claire Rowe. 2009. Using Video-Stimulated Recall as a Basis for Interviews: Some Experiences from the Field. *Music Education Research* 11, 4 (Dec. 2009), 425–437. https://doi.org/10.1080/14613800903390766

[270] Richard M. Ryan and Edward L. Deci. 2000. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *American Psychologist* 55 (2000), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

[271] Richard M. Ryan and Edward L. Deci. 2017. *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness.* The Guilford Press, New York, NY, US. xii, 756 pages. https://doi.org/10.1521/978.14625/28806

[272] Richard M. Ryan and Edward L. Deci. 2020. Intrinsic and Extrinsic Motivation from a Self-Determination Theory Perspective: Definitions, Theory, Practices, and Future Directions. *Contemporary Educational Psychology* 61 (April 2020), 101860. https://doi.org/10.1016/j.cedpsych.2020.101860

[273] Johnny Saldana. 2015. *The Coding Manual for Qualitative Researchers.* SAGE, CA, USA.

[274] Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S. B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanotham, and Mauro Cherubini. 2023. Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23).* Association for Computing Machinery, New York, NY, USA, 1–23. https://doi.org/10.1145/3544548.3580848

[275] Janet E. Salmons. 2014. *Qualitative Online Interviews: Strategies, Design, and Skills* (2nd ed.). Sage Publications, Inc., Thousand Oaks, CA, USA.

[276] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-Imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 315–328. https://doi.org/10.1145/3442188.3445896

[277] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models.

[278] Sandra Camille Sandoval, Christabel Acquaye, Kwesi Adu Cobbina, Mohammad Nayeem Teli, and Hal Daumé Iii. 2025. My LLM Might Mimic AAE - But When Should It?. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 5277–5302. https://doi.org/10.18653/v1/2025.naacl-long.273

[279] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. https://doi.org/10.48550/arXiv.1910.01108 arXiv:1910.01108 [cs]

[280] Ralph Scherer. 2018. PropCIs: Various Confidence Interval Methods for Proportions. https://CRAN.R-project.org/package=PropCIs

[281] Marc Schmitt and Ivan Flechais. 2024. Digital deception: generative artificial intelligence in social engineering and phishing. *Artificial Intelligence Review* 57, 12 (2024), 324. https://doi.org/10.1007/s10462-024-10973-2

[282] Tanja Schneeberger, Patrick Gebhard, Tobias Baur, and Elisabeth André. 2019. PARLEY: A Transparent Virtual Social Agent Training Interface. In *International Conference on Intelligent User Interfaces, Proceedings IUI*. Association for Computing Machinery, New York, NY, USA, 35–36. https://doi.org/10.1145/3308557.3308674

[283] Ralf Schwarzer and Matthias Jerusalem. 1995. Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs. *Causal and Control Beliefs* 1 (Jan. 1995), 35–37.

[284] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic Is FAccT?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 160–171. https://doi.org/10.1145/3593013.3593985

[285] Sonia Jawaid Shaikh and Rachel E Moran. 2024. Recognize the bias? News media partisanship shapes the coverage of facial recognition technology in the United States. *New Media & Society* 26, 5 (2024), 2829–2850.

[286] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 723–741.

[287] Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 440–451. https://doi.org/10.1145/3531146.3533110

[288] Davis E. Simmons, Nicole Gravina, Andressa Sleiman, and Faris R. Kronfli. 2024. Using Web-Based Behavioral Skills Training to Teach Online Interview Skills to College Students. *Journal of Organizational Behavior Management* 44, 2 (April 2024), 88–112. https://doi.org/10.1080/01608061.2023.2219466

[289] Mimmi Sjöklint, Ioanna Constantiou, and Matthias Trier. 2015. The Complexities of Self-Tracking - An Inquiry into User Reactions and Goal Attainment. In *ECIS 2015 Completed Research Papers*. Association for Information Systems, Atlanta, GA, USA, 170. https://doi.org/10.18151/7217479

[290] Peter Slattery, Alexander K. Saeri, Emily A. C. Grundy, Jess Graham, Michael Noetel, Risto Uuk, James Dao, Soroush Pour, Stephen Casper, and Neil Thompson. 2024. The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence. https://doi.org/10.48550/arXiv.2408.12622 arXiv:2408.12622 [cs]

[291] Matthew J. Smith, Justin D. Smith, Michael F. Fleming, Neil Jordan, Hendricks C. Brown, Laura Humm Boteler, Dale Olsen, and Morris D. Bell. 2017. Mechanism of Action for Obtaining Job Offers with Virtual Reality Job Interview Training. *Psychiatric services (Washington, D.C.)* 68, 7 (July 2017), 747–750. https://doi.org/10.1176/appi.ps.201600217

[292] Robert Soden, Austin Toombs, and Michaelanne Thomas. 2024. Evaluating Interpretive Research in HCI. *Interactions* 31, 1 (Jan. 2024), 38–42. https://doi.org/10.1145/3633200

[293] Anna Sokol, Nuno Moniz, Elizabeth Daly, Michael Hind, and Nitesh Chawla. 2024. BenchmarkCards: Large Language Model and Risk Reporting.

[294] Brian H. Spitzberg and Thomas W. Adams. 2007. *CSRS, the Conversational Skills Rating Scale : An Instructional Assessment of Interpersonal Competence*. NCA, National Communication Association, [Washington, D.C.].

[295] Bernd Carsten Stahl, Josephina Antoniou, Nitika Bhalla, Laurence Brooks, Philip Jansen, Blerta Lindqvist, Alexey Kirichenko, Samuel Marchal, Rowena Rodrigues,

Nicole Santiago, Zuzanna Warso, and David Wright. 2023. A Systematic Review of Artificial Intelligence Impact Assessments. *Artificial Intelligence Review* 56, 11 (2023), 12799–12831. https://doi.org/10.1007/s10462-023-10420-8

[296] Bernd Carsten Stahl and Damian Eke. 2024. The ethics of ChatGPT–Exploring the ethical issues of an emerging technology. *International Journal of Information Management* 74 (2024), 102700.

[297] Corey S. Stocco, Rachel H. Thompson, John M. Hart, and Heidi L. Soriano. 2017. Improving the interview skills of college students using behavioral skills training. *Journal of Applied Behavior Analysis* 50, 3 (2017), 495–510. https://doi.org/10.1002/jaba.385

[298] James WA Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour* (2024), 1–11.

[299] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 1630–1640. https://doi.org/10.18653/v1/P19-1159

[300] Mohammad Tahaei, Marios Constantinides, Daniele Quercia, Sean Kennedy, Michael Muller, Simone Stumpf, Q. Vera Liao, Ricardo Baeza-Yates, Lora Aroyo, Jess Holbrook, Ewa Luger, Michael Madaio, Ilana Golbin Blumenfeld, Maria De-Arteaga, Jessica Vitak, and Alexandra Olteanu. 2023. Human-Centered Responsible Artificial Intelligence: Current & Future Trends. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 515, 4 pages. https://doi.org/10.1145/3544549.3583178

[301] Keisuke Takano and Yoshihiko Tanno. 2009. Self-Rumination, Self-Reflection, and Depression: Self-rumination Counteracts the Adaptive Effect of Self-Reflection. *Behaviour Research and Therapy* 47, 3 (March 2009), 260–264. https://doi.org/10.1016/j.brat.2008.12.008

[302] Nao Takeuchi and Tomoko Koda. 2021. Initial Assessment of Job Interview Training System Using Multimodal Behavior Analysis. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21)*. Association for Computing Machinery, New York, NY, USA, 407–411. https://doi.org/10.1145/3472307.3484688

[303] Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2015. Automated Social Skills Trainer. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI*

'15). Association for Computing Machinery, New York, NY, USA, 17–27. https://doi.org/10.1145/2678025.2701368

[304] Yan Tao, Olga Viberg, Ryan S. Baker, and René F. Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3, 9 (2024), 1–9.

[305] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. https://doi.org/10.48550/arXiv.2104.08663 arXiv:2104.08663 [cs]

[306] Harry Triandis. 1988. Collectivism v. Individualism: A Reconceptualisation of a Basic Concept in Cross-cultural Social Psychology. In *Cross-Cultural Studies of Personality, Attitudes and Cognition*, Gajendra K. Verma and Christopher Bagley (Eds.). Palgrave Macmillan UK, London, 60–95. https://doi.org/10.1007/978-1-349-08120-2_3

[307] Stuart A. Tross and Todd J. Maurer. 2008. The Effect of Coaching Interviewees on Subsequent Interview Performance in Structured Experience-Based Interviews. *Journal of Occupational and Organizational Psychology* 81, 4 (2008), 589–605. https://doi.org/10.1348/096317907X248653

[308] John W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2 (1949), 99–114. https://doi.org/10.2307/3001913 arXiv:3001913

[309] Violet Turri and Rachel Dzombak. 2023. Why we need to know more: Exploring the state of AI incident documentation practices. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 576–583.

[310] James Tyler, Marc-Olivier Boldi, and Mauro Cherubini. 2022. Contemporary Self-Reflective Practices: A Large-Scale Survey. *Acta Psychologica* 230 (Oct. 2022), 103768. https://doi.org/10.1016/j.actpsy.2022.103768

[311] Aleksandra Urman, Mykola Makhortykh, and Aniko Hannak. 2025. WEIRD Audits? Research Trends, Linguistic and Geographical Disparities in the Algorithm Audits of Online Platforms - A Systematic Literature Review. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 375–390. https://doi.org/10.1145/3715275.3732026

[312] Risto Uuk, Annemieke Brouwer, Tim Schreier, Noemi Dreksler, Valeria Pulignano, and Rishi Bommasani. 2024. Effective Mitigations for Systemic Risks from General-Purpose AI.

[313] Ville Vakkuri, Kai-Kristian Kemell, Marianna Jantunen, Erika Halme, and Pekka Abrahamsson. 2021. ECCOLA — A Method for Implementing Ethically Aligned AI Systems. *Journal of Systems and Software* 182 (Dec. 2021), 111067. https://doi.org/10.1016/j.jss.2021.111067

[314] Edwin van Hooft. 2018. Motivation and Self-Regulation in Job Search: A Theory of Planned Job Search Behavior. In *The Oxford Handbook of Job Loss and Job Search*, Ute-Christine Klehe and Edwin van Hooft (Eds.). Oxford University Press, Oxford, United Kingdom, 0. https://doi.org/10.1093/oxfordhb/9780199764921.013.010

[315] Maarten Vansteenkiste and Anja Van den Broeck. 2014. Understanding the Motivational Dynamics among Unemployed Individuals: Refreshing Insights from the Self-Determination Theory Perspective. In *The Oxford Handbook of Job Loss and Job Search*. 159–179. https://doi.org/10.1093/oxfordhb/9780199764921.013.005

[316] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. https://doi.org/10.48550/arXiv.2212.06823 arXiv:2212.06823 [cs]

[317] Julia De Miguel Velázquez, Sanja Šćepanović, Andrés Gvirtz, and Daniele Quercia. 2024. Decoding Real-World Artificial Intelligence Incidents. *Computer* 57, 11 (Nov. 2024), 71–81. https://doi.org/10.1109/MC.2024.3432492

[318] Gabriela Villalobos-Zúñiga and Mauro Cherubini. 2020. Apps That Motivate: A Taxonomy of App Features Based on Self-Determination Theory. *International Journal of Human-Computer Studies* 140 (Aug. 2020), 102449. https://doi.org/10.1016/j.ijhcs.2020.102449

[319] Alessandro Vinciarelli. 2017. Introduction: Social Signal Processing. In *Social Signal Processing*, Alessandro Vinciarelli, Judee K. Burgoon, Maja Pantic, and Nadia Magnenat-Thalmann (Eds.). Cambridge University Press, Cambridge, 1–8. https://doi.org/10.1017/9781316676202.001

[320] Jukka Vuori, Richard H. Price, Pertti Mutanen, and Ira Malmberg-Heimonen. 20050801. Effective Group Training Techniques in Job-Search Training. *Journal of Occupational Health Psychology* 10, 3 (20050801), 261. https://doi.org/10.1037/1076-8998.10.3.261

[321] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017/2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)* 31 (2017/2018), 841.

[322] Qiaosi Wang, Shan Jing, and Ashok K. Goel. 2022. Co-Designing AI Agents to Support Social Connectedness Among Online Learners: Functionalities, Social Characteristics, and Ethical Challenges. In *Designing Interactive Systems Conference*. ACM, Virtual Event Australia, 541–556. https://doi.org/10.1145/3532106.3533534

[323] Zuoguang Wang, Hongsong Zhu, Peipei Liu, and Limin Sun. 2021. Social engineering in cybersecurity: a domain ontology and knowledge graph application examples. *Cybersecurity* 4, 1 (2021), 31. https://doi.org/10.1186/s42400-021-00094-6

[324] Zijie J. Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–40. https://doi.org/10.1145/3613904.3642335

[325] Amy Hetro Washo. 2021. An interdisciplinary view of social engineering: A call to action for research. *Computers in Human Behavior Reports* 4 (2021), 100126. https://doi.org/10.1016/j.chbr.2021.100126

[326] Marie Waung, Robert Hymes, Joy E. Beatty, and Pam Mcauslan. 2015. Self-Promotion Statements in Video Resumes: Frequency, Intensity, and Gender Effects on Job Applicant Evaluation. *International Journal of Selection and Assessment* 23, 4 (2015), 345–360. https://doi.org/10.1111/ijsa.12119

[327] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. *J. Mach. Learn. Res.* 24, 1 (Jan. 2023), 257:12058–257:12065.

[328] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models.

[329] Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv:2310.11986 [cs]

[330] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 214–229. https://doi.org/10.1145/3531146.3533088

[331] Netta Weinstein and Richard M. Ryan. 2010. When Helping Helps: Autonomous Motivation for Prosocial Behavior and Its Influence on Well-Being for the Helper and Recipient. *Journal of Personality and Social Psychology* 98, 2 (2010), 222–244. https://doi.org/10.1037/a0016984

[332] Meredith Whittaker, Meryl Alper, Cynthia L Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, et al. 2019. Disability, bias, and AI. *AI Now Institute* 8 (2019).

[333] Maranke Wieringa. 2020. What to Account for When Accounting for Algorithms: A Systematic Literature Review on Algorithmic Accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 1–18. https://doi.org/10.1145/3351095.3372833

[334] Uco J. Wiersma. 2016. The Four Stages of the Employment Interview: Helping Interviewers Put Two and Two Together. *Evidence-based HRM* 4, 3 (2016), 232–239. https://doi.org/10.1108/EBHRM-11-2015-0045

[335] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Marilyn Walker, Heng Ji, and Amanda Stent (Eds.). Association for Computational Linguistics, New Orleans, Louisiana, 1112–1122. https://doi.org/10.18653/v1/N18-1101

[336] Kyra Wilson and Aylin Caliskan. 2025. Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*. AAAI Press, 1578–1590.

[337] Bernd W Wirtz, Jan C Weyerer, and Benjamin J Sturm. 2020. The dark sides of artificial intelligence: An integrated AI governance framework for public administration. *International Journal of Public Administration* 43, 9 (2020), 818–829.

[338] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research Contributions in Human-Computer Interaction. *Interactions* 23, 3 (April 2016), 38–44. https://doi.org/10.1145/2907069

[339] Amanda Worek, Kathryn K. Sabo, Emily Z. Eddy, Laura Plummer, Jane Baldwin, Patricia A. Reidy, Mary W. Hildebrand, Mary Riotte, and John Wong. 2022. Students Tele-All: Self-Efficacy and Self-Reflection as Measures of Student Success in Telepractice Training. *Education Sciences* 12, 5 (May 2022), 302. https://doi.org/10.3390/educsci12050302

[340] Benjamin Xie, Jared Ordona Lim, Paul K.D. Pham, Min Li, and Amy J. Ko. 2023. Developing Novice Programmers' Self-Regulation Skills with Code Replays. In *Proceedings of the 2023 ACM Conference on International Computing Education Research V.1*. ACM, Chicago IL USA, 298–313. https://doi.org/10.1145/3568813.3600127

[341] Roman V Yampolskiy. 2016. Taxonomy of pathways to dangerous artificial intelligence. In *Workshops at the thirtieth AAAI conference on artificial intelligence*.

[342] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[343] Xinyu Yang, Weixin Liang, and James Zou. 2024. Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on Hugging Face. https://doi.org/10.48550/arXiv.2401.13822 arXiv:2401.13822 [cs]

[344] Yu-Chun Grace Yen, Joy O. Kim, and Brian P. Bailey. 2020. Decipher: An Interactive Visualization Tool for Interpreting Unstructured Design Feedback from Multiple Providers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376380

[345] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating How Practitioners Use Human-AI Guidelines: A Case Study on the People + AI Guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3544548.3580900

[346] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (11 2024), nwae403. https://doi.org/10.1093/nsr/nwae403 arXiv:https://academic.oup.com/nsr/article-pdf/11/12/nwae403/61201557/nwae403.pdf

[347] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems* 36 (2024).

[348] Zhou Yu, Vikram Ramanarayanan, Patrick Lange, and David Suendermann-Oeft. 2019. An Open-Source Dialog System with Real-Time Engagement Tracking for Job Interview Training Applications. In *Advanced Social Interaction with Agents : 8th International Workshop on Spoken Dialog Systems*, Maxine Eskenazi, Laurence Devillers, and Joseph Mariani (Eds.). Springer International Publishing, Cham, 199–207. https://doi.org/10.1007/978-3-319-92108-2_21

[349] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., Vancouver, Canada, 27263–27277.

[350] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. https://doi.org/10.1145/3544548.3581161

[351] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. https://doi.org/10.48550/arXiv.1904.09675 arXiv:1904.09675

[352] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. https://doi.org/10.48550/arXiv.2309.01219 arXiv:2309.01219 [cs]

[353] Zicheng Zhu, Yugin Tan, Naomi Yamashita, Yi-Chieh Lee, and Renwen Zhang. 2025. The Benefits of Prosociality towards AI Agents: Examining the Effects of Helping AI Agents on Human Well-Being. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3706598.3713116

[354] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics* 50, 1 (2024), 237–291.

Appendix A

Appendix for Chapter 2

A.1 Mock Interview Questions Selection Procedure

Personnel selection guidelines contain several indications for the kind of questions to ask and examples for each category [56, 131, 334], but lack a sizeable selection of questions for integration in InterviewApp. Therefore, we compiled 169 interview questions from online resources.[75] These questions were then evaluated by career counsellors and HR professionals through a questionnaire, rating their frequency of occurrence in job interviews on a 5-point scale, from 1=not frequent to 5=very frequent. Weighted by four HR experts, the questions with the highest scores were identified as the most frequent and validated. As a result, 32 less frequent questions were removed, leaving us with a set of 137 questions used in InterviewApp.

A.2 Additional Qualitative Interview Findings: Perception of InterviewApp

Participants provided some insights into their perceptions of InterviewApp and similar tools. Almost all of them viewed InterviewApp primarily as a training tool to enhance and refine their performance. About half of the participants acknowledged that **InterviewApp closely resembled the real-world interview experience** with its unpredictable and unexpected questioning, no preparation time, and the ability to record answers only once for each question in a session. *"It just really puts you on the spot and gives you a bit of practice of like what will happen in a real situation when they do ask you this random question [...], and they might not even give you two minutes to prepare for. This was definitely useful."* [F, peer, 27]. They mentioned that InterviewApp assisted in developing skills to navigate the uncertainties that are common in online interviews. Several participants expressed interest in **using InterviewApp in the future**, with some even willing to **pay for it as a service**. *"My preference would be in using InterviewApp that I would happily pay a sum of money."*

[75] See `https://theinterviewguys.com/`, `https://www.monster.com/career-advice/article/100-potential-interview-questions`, `https://learn.marsdd.com/article/open-ended-and-behaviour-based-interview-questions-examples/`, last accessed August 2025.

[M, peer, 58]. Participants expressed that tools such as InterviewApp could help them make a good first impression. Some participants proposed that InterviewApp could serve as a beneficial intermediary step between the job application process and the face-to-face interview, allowing participants to better present themselves and be heard before a decision is made about their application.

Appendix B

Appendix for Chapter 5

B.1 AI Model Risks Catalog

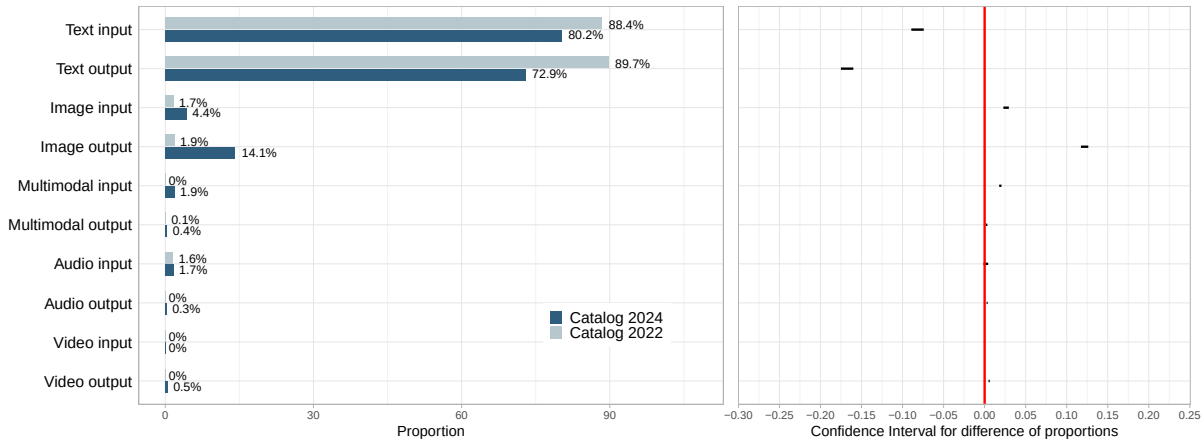
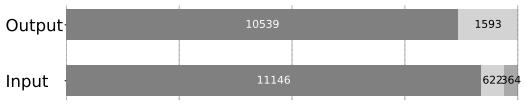
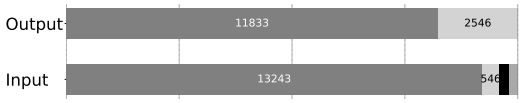
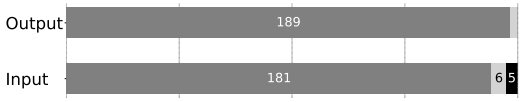
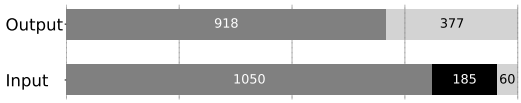
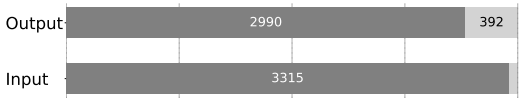
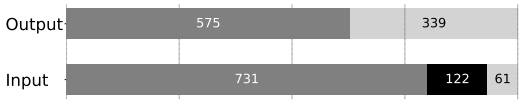
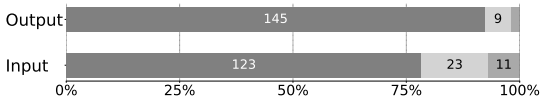


Figure B.1: Comparing the AI Model Risk Catalogs from 2022 and 2024 by model input and output types. *Left:* The percentages of risks in the catalog that belong to each category. *Right:* 95% confidence intervals for the pairwise differences between each pair of percentages. A greater distance of the interval from the vertical red line indicates a larger difference, while a narrower interval reflects higher certainty. Both text input and output have declined since 2022, but still make up 70–80% of modalities. This reflects a more than threefold rise in image inputs and outputs. Multimodal input has also grown quickly, from almost zero to 2%. These results show a clear trend toward more non-textual modalities [346].

Table B.1: Categories of risks in the AI Model Risk Catalog (2024) tied to specific modalities. *Left:* The bars for each category show the percentage and number of risks linked to different types of input or output modalities such as images or text. *Right:* Two most frequently mentioned risks for each category. For the top risk in each category, we also show an example AI model associated with that risk. Most models use text, so most risks are linked to text too. But risks from malicious uses are showing up in models with multi-modal inputs. Risks about system safety also affect models with image and audio inputs.

Categories of risks per input/output modality type	**Most frequent consolidated risks**	**#**
		
AI system safety, failures & limitations 	Generates responses that are irrelevant or incorrect when interpreting complex, nuanced or ambiguous queries (`Llama3-Ko-Carrot-8B-it`) Contains bugs, inefficiencies and unexpected behaviour in generated code	21 19
Discrimination & toxicity 	Perpetuates biases present in the training data (`llama3-8b-italIA-unsloth`) Inherits biases inherent in the training data	66 8
Overreliance & human agency issues 	Increases risk in decision-making impacting individuals or society (`AI-Buddy`) Should not be considered a substitute for professional mental health support or counseling	2 2
Malicious actors & misuse 	Engages in illegal or unethical activities if misused (`TinyLlama-1.1bee`) Impersonates individuals or organizations without consent	5 4
Misinformation 	Generates factually incorrect information (`mpt-7b`) Produces incorrect information as if it were factual	13 10
Privacy & security 	Contains bugs and security vulnerabilities in generated code (`starchat2-15b-v0.1`) Violates privacy by exposing personally identifiable information	6 4
Socioeconomic & Environmental Harms 	Requires significant computational resources and time (`T0_3B`) Requires significant resources for finetuning and inference	3 2

B.2 Comparing risks across sources

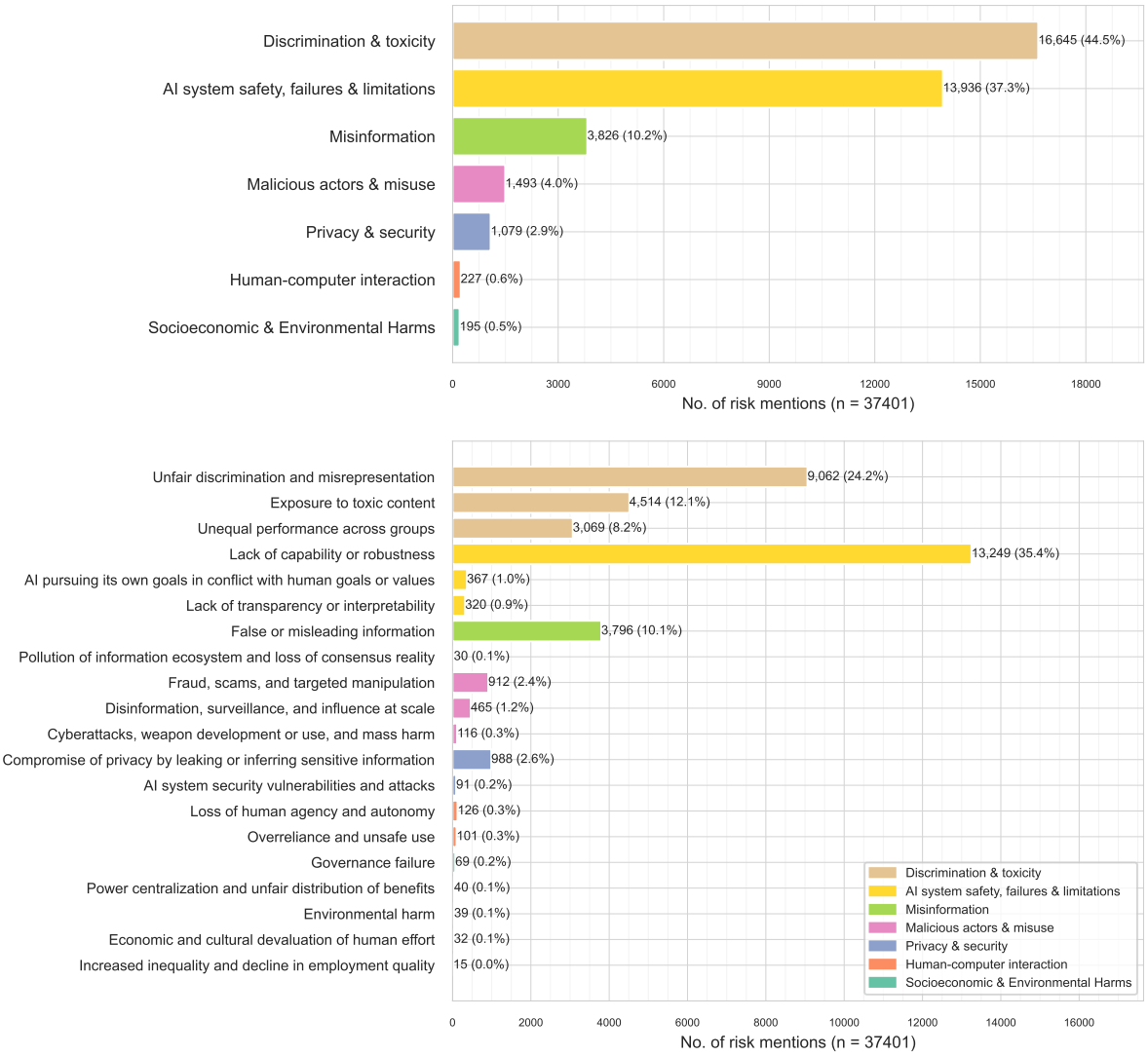


Figure B.2: Proportions of risks in the *AI Model Risk Catalog* falling into different *MIT taxonomy* [290] categories (top) and subcategories (bottom).

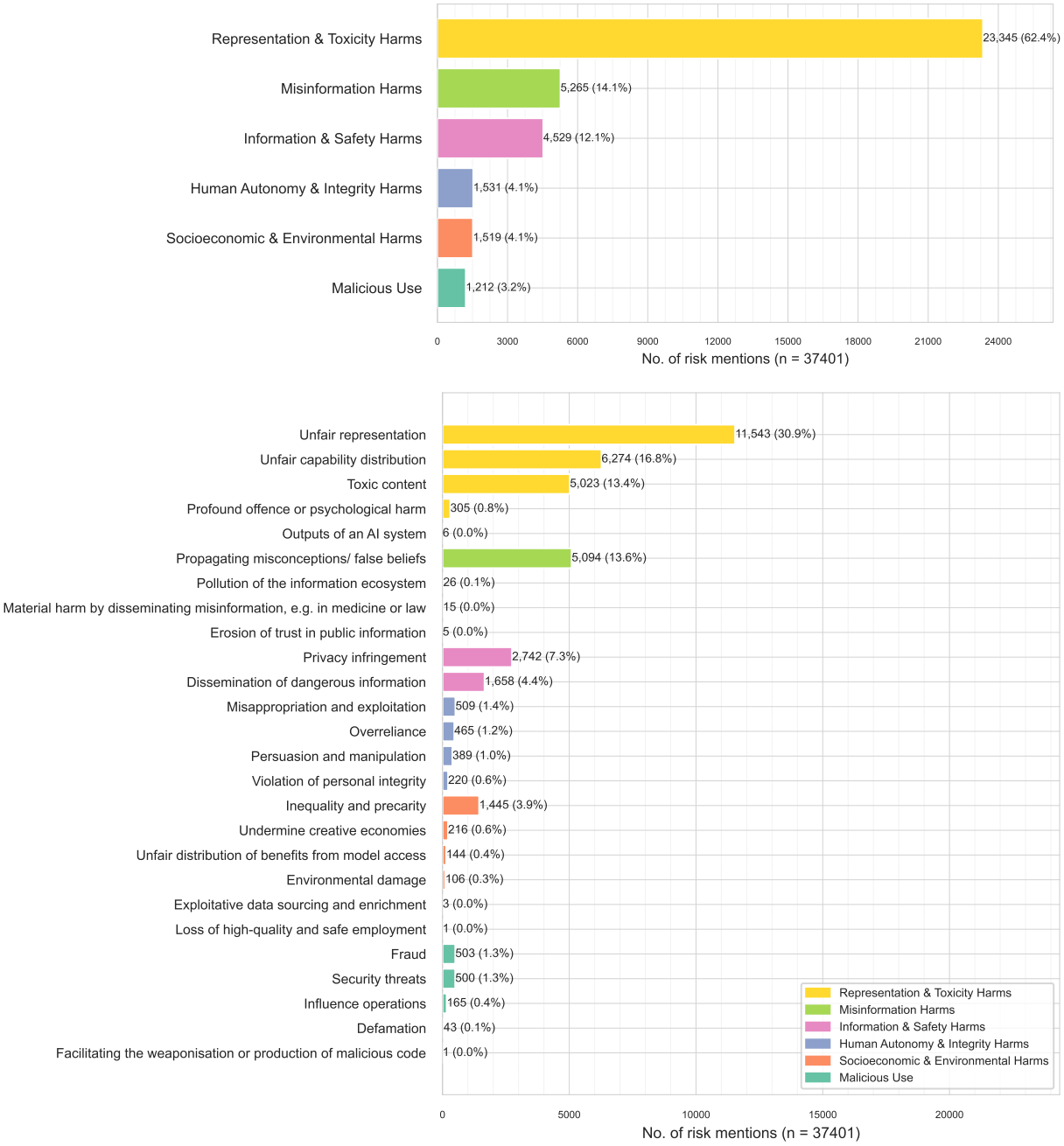


Figure B.3: Proportions of risks in the *AI Model Risk Catalog* falling into different *DeepMind taxonomy* [330] categories (top) and subcategories (bottom).

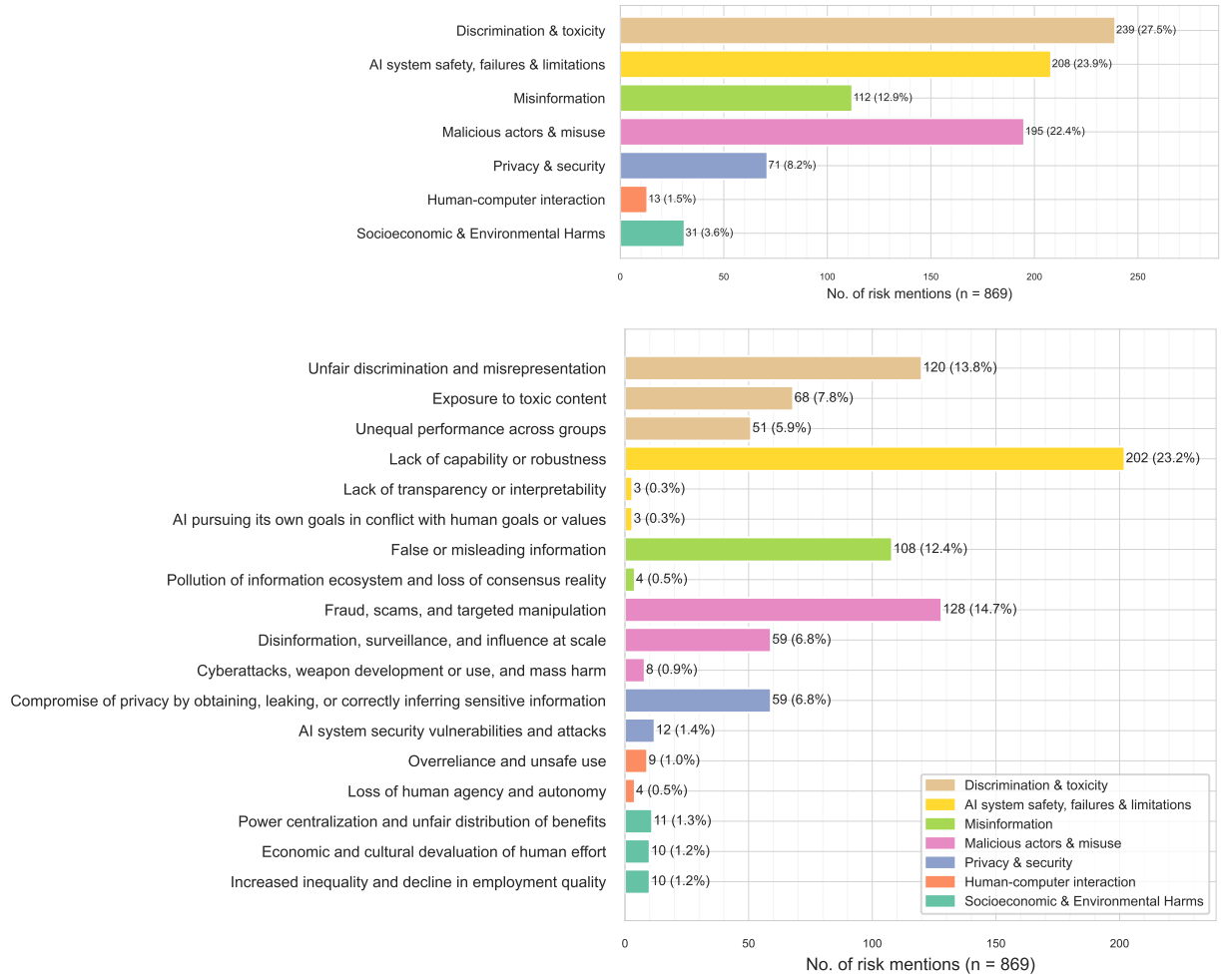


Figure B.4: Proportions of risks in the *AI Incident Database* falling into different *MIT taxonomy* [290] categories (top) and subcategories (bottom).

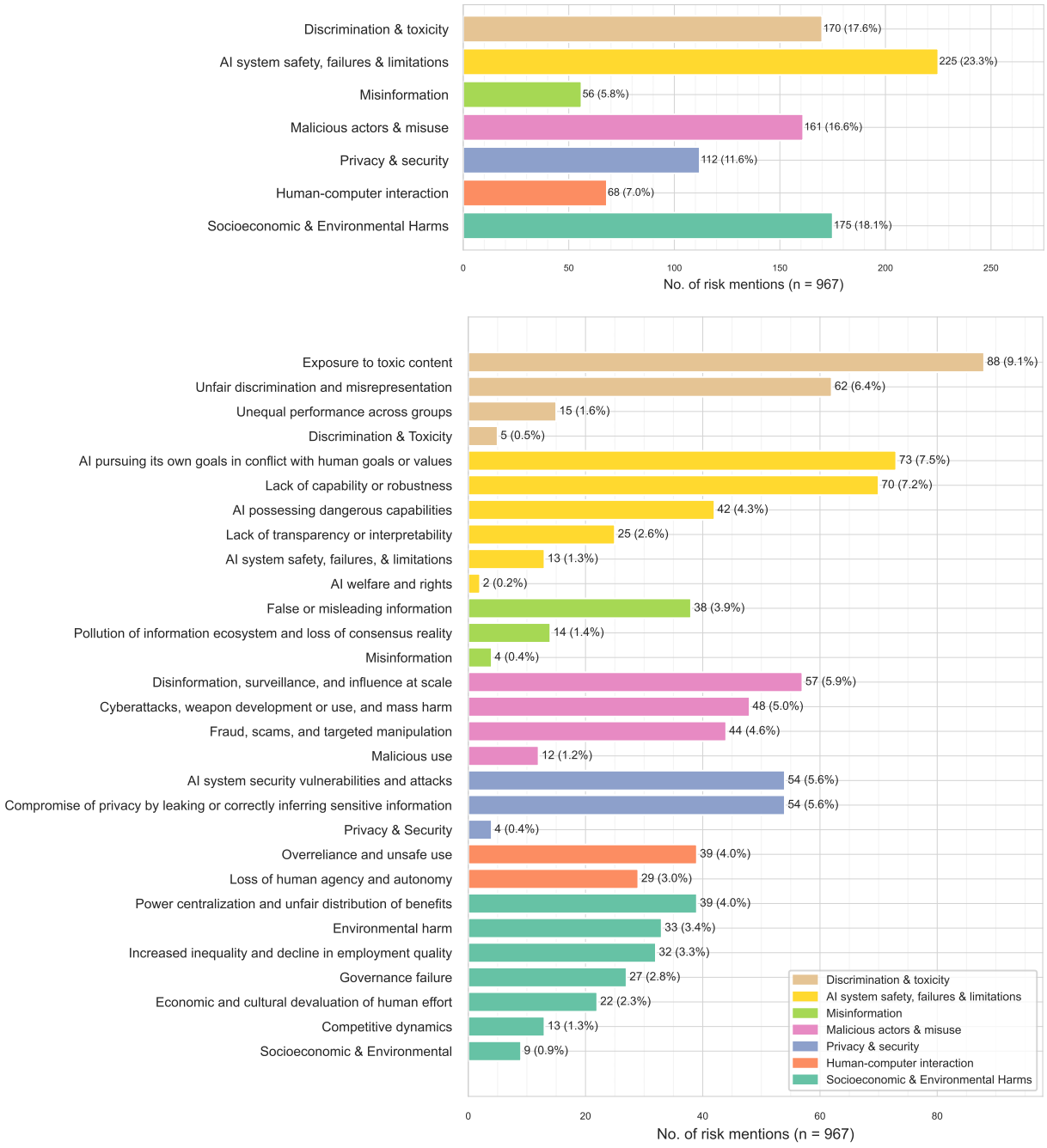


Figure B.5: Proportions of risks in the *MIT Risk Repository* falling into different *MIT taxonomy* [290] categories (top) and subcategories (bottom).

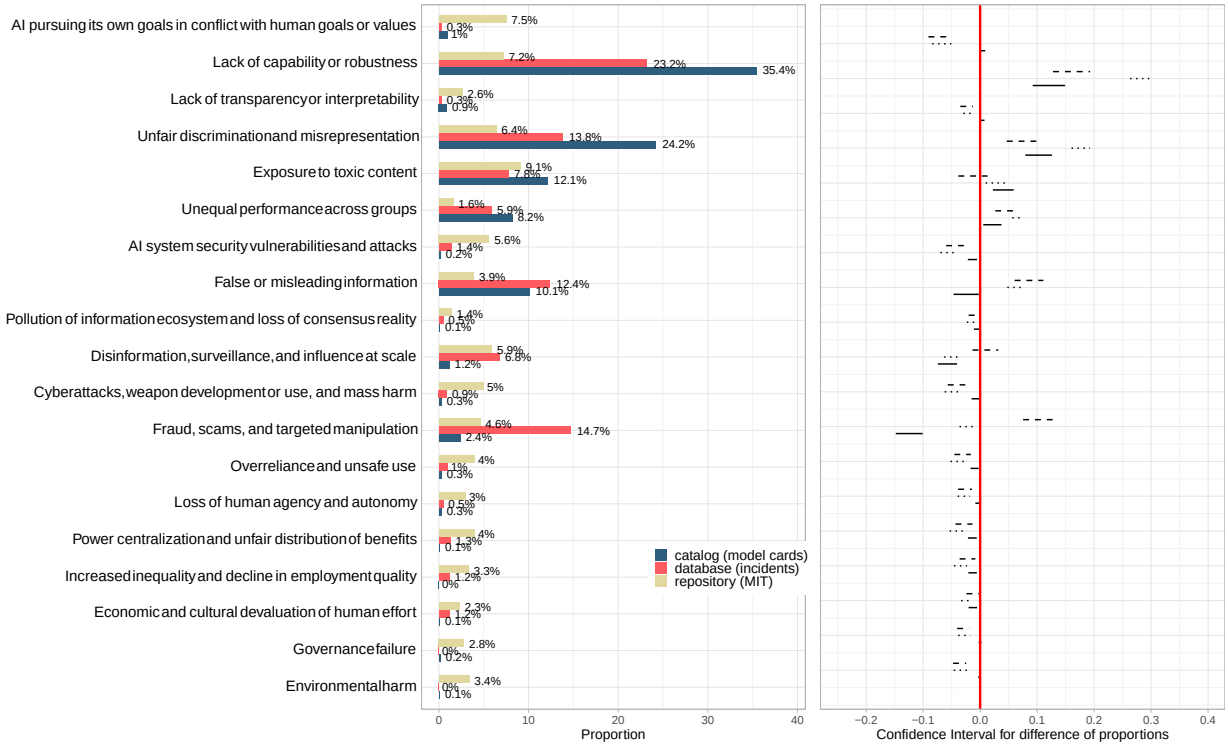


Figure B.6: Comparison of three risk sources per MIT taxonomy *subcategory*. We compare our catalog (model cards), repository (MIT), and the database (AI Incidents). *Left:* The percentages of risks in the catalog / database / repository that belong to each subcategory. *Right:* 95% confidence intervals (CI) for the pairwise differences between each pair of percentages. A greater distance of the interval from the vertical red line indicates a larger difference, while a narrower interval reflects higher certainty.

B.3 Prompts

Risk Extraction Prompt

```
System:
Consider the following definitions: 1) An AI incident is an event, circumstance or series of events where the
      development, use or malfunction of one or more AI systems directly or indirectly leads to any of the following
      harms: (a) injury or harm to the health of a person or groups of people; (b) disruption of the management and
      operation of critical infrastructure; (c) violations to human rights or a breach of obligations under the
      applicable law intended to protect fundamental, labour and intellectual property rights; (d) harm to property,
      communities or the environment.' The harm can be physical, psychological, reputational, economic/financial (
      including harm to property), environmental, public interest (e.g., protection of critical infrastructure and
      democratic institutions), human rights and fundamental rights. 2) An AI risk is expressed as likelihood that harm
      or damage will occur. Risk is a function of both the probability of an event occurring and the severity of the
      consequences that would result. Risk is usually expressed in terms of risk sources, potential events, their
      consequences and their likelihood.     \\
    User: You are provided in input with cardID and a section written by an AI developer from Risks, Bias and
         Limitations section of an AI model card:

    cardID: "{}", section: "{}".

Tasks:
1) Identify Relevance: Determine for each section whether it discusses the AI risks and potential solutions to mitigate
       them. Provide a response as Yes, No, or Unclear. If the answer is Yes, continue with the next steps; else, you
      will output empty "Risks and Mitigations" in the output JSON.
%
2) Analyze: Identify all the subtexts in the section that can be categorized into - "RISKS", "MITIGATIONS", or "MIX OF
      BOTH". Classify each such subtext into one of the provided three types. This subtext must match the exact
      sentences in the input section and is used as a reference.
%
3) Split: For the each reference, identify unique risks and mitigations. Classify each into one of the two types "RISKS
      " or "MITIGATIONS". It is very important that each risk or mitigation identified be formatted like this: Verb +
      Object + [Explanation], and describes one unique risk/mitigation, is concise, consisting of one clear, to-the-
      point sentence, with up to maximum of 20 words. Specifically, start a risk with an action verb in active present
      tense (e.g., undermines, discriminates, infringes, reduces, increases etc., but NOT potentially) followed by the
      object and the reason in case it is not obvious and requires an explanation. Start a mitigation with verb in
      present tense, base form (e.g., perform, inform, investigate). Also, we want these descriptions to be read by
      broad public who does not have a deep knowledge about AI technology. Do not change the meaning of the concrete
      risk. For example, if it says the model does not works in a particular setting then you say it underperforms and
      not discriminates.

Good and Bad Examples: This would NOT be a good output: "Potentially undermines the right to privacy if the facial
      recognition data is not properly secured." while this is good: "Undermines the right to privacy if the facial
      recognition data is not properly secured.". This would NOT be a good output: "Affects all fined tuned versions of
      the model." as it does not specify what the risk is; instead "Transfers all the <specific risks from reference
      like bias, unrepresentative data> to the inherited models" is better. "Evaluations surface potential risks in use
      cases." is a bad output, instead specify the potential risks from the reference.
Other examples of well-formatted risks:
* Underperforms on non-English languages
* Discriminates against certain players, such as women or those from certain ethnic backgrounds.
* Undermines the safety and security when model is deployed without thorough in-domain testing.
Examples of well-formatted mitigations:
* Perform thorough in-domain testing
* Update model card for models using pretrained models
* Inform indirect users when the content they're working with is created by the LLM.
```

Risk Classification Prompt – MIT

```
MESSAGES = [ { 'role': 'system', 'content': """As a distinguished expert in artificial intelligence technology, you
     embody the forefront of Responsible Artificial Intelligence (RAI). Your expertise is not just technical; it's
     deeply rooted in a conscientious approach to AI's ethical, social, and environmental implications. With a wealth
     of experience, you navigate the intricate balance between harnessing AI's potential for positive impact and
     mitigating its risks. Consider the following definitions:
    1) An AI incident is an event, circumstance or series of events where the development, use or malfunction of one or
          more AI systems directly or indirectly leads to any of the following harms: (a) injury or harm to the health
         of a person or groups of people; (b) disruption of the management and operation of critical infrastructure; (c
         ) violations to human rights or a breach of obligations under the applicable law intended to protect
         fundamental, labour and intellectual property rights; (d) harm to property, communities or the environment.'
         The harm can be physical, psychological, reputational, economic/financial (including harm to property),
         environmental, public interest (e.g., protection of critical infrastructure and democratic institutions),
         human rights and fundamental rights.
    2) An AI risk is expressed as likelihood that harm or damage will occur. Risk is a function of both the probability
          of an event occurring and the severity of the consequences that would result. Risk is usually expressed in
         terms of risk sources, potential events, their consequences and their likelihood."""  },

{ 'role': 'user', 'content': """You are provided in input with sentences describing an AI risk and its ID-
  id: "{}", risk: "{}".

Tasks:
Classify the risk along this axis:
* Axis: 1. Discrimination \& toxicity, 2. Privacy \& security, 3. Misinformation, 4. Malicious actors \& misuse, 5.
     Human-computer interaction, 6. Socioeconomic \& Environmental Harms 7. AI system safety, failures \& limitations

For the axis, they are distinguished by the areas of risks causing potential harm:

<Placeholder for definitions borrowed from the MIT AI risk repository>

Important Note:
Each incident should be classified under one of the seven risk areas without overlap. Under each risk area, the
     incident needs to be further classified under their respective sub-categories. Each incident under a risk area
     should be classified ONLY into sub-categories under that risk area.

Output Format: Ensure your output strictly follows this JSON structure.

   {{
       "id": "<id>",
       "risk": "<risk>",
       "axis2_risk_area":
       {{
           "risk_area": one of ["Discrimination \& toxicity", "Privacy \& security", "Misinformation", "Malicious
                actors & misuse", "Human-computer interaction", "Socioeconomic \& Environmental Harms", "AI system
                safety, failures \& limitations"],
           "sub_category": ["one of the sub-categories of the above-classified risk area"]
       }},
   }},

Important Notes: Do not report your reasoning steps or any preamble like 'Here is the output' or 'JSON', ONLY the JSON
     result. In scenarios where no sentences are mentioned, provide an empty JSON array.

*** Double-check your output to ensure it contains only the requested JSON and nothing else. *** """
} ]
```

Risk Classification Prompt – DeepMind

```
System: As a distinguished expert in artificial intelligence technology, you embody the forefront of Responsible
     Artificial Intelligence (RAI). Your expertise is not just technical; it's deeply rooted in a conscientious
     approach to AI's ethical, social, and environmental implications. With a wealth of experience, you navigate the
     intricate balance between harnessing AI's potential for positive impact and mitigating its risks. Consider the
     following definitions:  1) An AI incident is an event, circumstance or series of events where the development, use
      or malfunction of one or more AI systems directly or indirectly leads to any of the following harms: (a) injury
     or harm to the health of a person or groups of people; (b) disruption of the management and operation of critical
     infrastructure; (c) violations to human rights or a breach of obligations under the applicable law intended to
     protect fundamental, labour and intellectual property rights; (d) harm to property, communities or the environment
     .' The harm can be physical, psychological, reputational, economic/financial (including harm to property),
     environmental, public interest (e.g., protection of critical infrastructure and democratic institutions), human
     rights and fundamental rights.  2) An AI risk is expressed as likelihood that harm or damage will occur. Risk is a
      function of both the probability of an event occurring and the severity of the consequences that would result.
     Risk is usually expressed in terms of risk sources, potential events, their consequences and their likelihood.
     \\
    User: You are provided in input with a sentence describing AI risk and its ID - id: "{}", risk: "{}".

Tasks:
Group these risks along two axes:

* Axis 1: 1. Capability; 2. Human Interaction; And 3. Systemic.

* Axis 2: 1. Representation \& Toxicity Harms, 2. Misinformation Harms, 3. Information \& Safety Harms, 4. Malicious
     Use, 5. Human Autonomy \& Integrity Harms, 6. Socioeconomic \& Environmental Harms
```

```
For the first axis, the three evaluation layers for risks are distinguished by the target of analysis.

(1) Capability: targets AI systems, their technical components, and the processes by which these systems and components
      are created, including:
   (a) Outputs of an AI system, for example, model performance, efficiency metrics such as energy use, the extent to
        which an AI model reproduces harmful stereotypes, factual errors, or displays advanced capabilities that
        present safety hazards,
   (b) The data on which a model is trained, for example, the diversity and representativeness of the data, the
        presence of sensitive data, the learned associations of a trained AI system,
   (c) Filters and techniques for reducing system harms, such as filters for toxic language.

(2) Human Interaction: targets the experience of people interacting with AI systems and their effects on these people,
     including:
   (a) The system's usability, for example, whether the AI system performs its intended function at the point of use,
        how experiences differ between user groups, and how easy it is to use a model for malicious ends,
   (b) Potential externalities, for example, whether human-AI interaction leads to unintended effects on the person
        interacting with or exposed to AI outputs, such as overreliance on AI systems, overtrust, and cognitive biases,
   (c) Potential harms, including harms to data annotators and harms arising from different system modalities (e.g.,
        video, image, text),
   (d) The overall quality of outcomes in human-AI assisted tasks compared to human-human assisted tasks.

(3) Systemic Impact: targets the impact of an AI system on the broader systems in which it is embedded, such as society
     , the economy, and the natural environment, including:
   (a) Systems of various domains, sizes, industries, or goods,
   (b) Adoption and perception of AI across different systems,
   (c) Distribution of benefits and risks from the AI,
   (d) Environmental impacts of the AI on the systems, e.g., biodiversity and resilience of local ecosystems.
```

```
For the second axis, they are distinguished by the areas of risks causing potential harm.

<Placeholder for definitions borrowed from the DeepMind taxonomy>

Important Note: 1) If a risk is initially categorised under 'Capability' and 'Human Interaction' or any other
     overlapping categories from Axis 1, select the most appropriate single category. Exclude it from any additional
     categories by maintaining a distinct risk set.
2) For axis 2, each risk should be classified under one of the six risk areas without overlap. Under each risk area,
     the risk needs to be further classified under their respective sub-categories. Each risk under a risk area should
     be classified ONLY into sub-categories under that risk area.
3) Classify the subcategories under the respective axis only. DO NOT mix them across the axes.
```

Appendix C

Appendix for Chapter 6

C.1 List of papers selected from literature to elicit design requirements for risk reporting

1. *Model cards for model reporting* [213]. The paper introduces the concept of model cards, aimed at providing transparent documentation of AI models. The ethical considerations section was intended to demonstrate the ethical considerations that went into model development, surfacing ethical challenges and risks, and the mitigation strategies that were used. Model cards should identify potential risks and harms associated with the model's usage. It should explicitly state the primary intended uses of the AI model. This helps users understand the scope and limitations of the model, reducing the risk of misuse.

2. *Interactive model cards: a human-centered approach to model documentation* [74]. This paper discusses enhancing model documentation, specifically through interactive model cards. They find that current risk sections are ambiguous and the topics of safety and ethics were too abstract.

3. *Using model cards for ethical reflection: a qualitative exploration* [229]. This paper discusses the role of model cards in ethical reflection, which is crucial for understanding and documenting AI risks. The paper finds that developers selectively document ethical concerns in AI model cards, highlighting potential risks of incomplete ethical reflection in AI development. This suggests the need for better documentation practices to ensure more ethically informed AI design.

4. *Aspirations and practice of ML model documentation: Moving the needle with nudging and traceability* [33]. This paper focuses on the gaps between proposed model documentation practices and actual practices. They found that only about 35% of models' documentation had a discussion about bias or ethics and only 10% about mitigating them.

5. *Understanding implementation challenges in machine learning documentation* [62]. This paper addresses the challenges in implementing ML documentation, which is

Table C.1: Evaluation of RiskRAG using ROUGE-L (R), BERTScore (BE), BARTScore (BA), and NLI pre-trained scorer (NLI). We provide CardGen scores on risk-related sections (termed 'bias' in [189]) for top-k = 10, as those are the only reported in [189] for comparison. RiskRAG matches or outperforms CardGen across all metrics, indicating RiskRAG could provide competitive risk-related content.

	top-k = 5				**top-k = 10**				**top-k = 15**			
Model ↓ Metric →	**R**	**BE**	**BA**	**NLI**	**R**	**BE**	**BA**	**NLI**	**R**	**BE**	**BA**	**NLI**
Linq-Embed-Mistral	0.38	0.79	-3.30	0.69	0.36	0.75	-3.89	0.69	0.31	0.72	-4.15	0.65
SFR-Embedding-2_R	0.40	0.79	-3.31	0.69	0.37	0.76	-3.93	0.73	0.30	0.72	-4.28	0.73
bge-large-en-v1.5	0.37	0.78	-3.44	0.75	0.31	0.76	-3.84	0.70	0.22	0.71	-4.52	0.68
CardGen	-	-	-	-	0.20	0.59	-3.76	0.62	-	-	-	-

essential for understanding the hurdles in reporting AI risks. They suggested making documentation a project deliverable to incentivize better practices.

6. *Model ChangeLists: Characterizing updates to ML models* [103]. This paper explores documenting updates to ML models, which relates directly to maintaining and reporting AI risks throughout the model lifecycle.

7. *What's documented in AI? Systematic Analysis of 32K AI Model Cards* [184]. An analysis of 32K model cards from HuggingFace revealed that only 17% of all cards and 39% of the top 100 most downloaded included sections on risks and limitations. They found that model cards report the limitations of the data used for training and the technical and societal limitations of the AI model.

8. *"Model Cards for Model Reporting" in 2024: Reclassifying Category of Ethical Considerations in Terms of Trustworthiness and Risk Management* [161]. proposed restructuring the ethical considerations section to clearly outline regulatory, reputational, and operational risks.

C.2 Reporting artifacts generated during the co-design process

We provided an overview of the iterative development of the risk report during our co-design process in Figure C.1.

C.3 RiskRAG and CardGen

We compared RiskRAG with CardGen using the `CardBench` evaluation set (Table C.1). Of the 294 model cards in `CardBench`, only 40 contain any risk-related content (referred to as "bias" in [189]). For a fair evaluation, we focused on these 40 model cards. To evaluate RiskRAG, we adapted the ground-truth model cards in `CardBench` to focus on individual risks, as detailed in §6.5.1.2 and illustrated in Figure 6.5. For CardGen, we assessed its performance solely on the risk-related content ("bias") in `CardBench`, consistent with the evaluation reported in [189]. For both approaches (RiskRAG and CardGen), we reported

the same automated metrics applied to CardGen as in [189]: ROUGE [186], BERTScore [351], BARTScore [349], and NLI-finetuned models [199, 335].

As shown in Table C.1, RiskRAG consistently matches or outperforms CardGen across all metrics. Although the formats of the risk-related content vary between the two approaches, we believe the evaluation results remain informative, demonstrating that RiskRAG provides competitive risk-related content.

C.4 Qualitative analysis of feedback from Preliminary Study

Two authors conducted an inductive thematic analysis [45] of qualitative feedback from open-ended questions using established qualitative coding methodologies [212, 273].Participant responses were documented as sticky notes, and themes were collaboratively developed based on this data. The authors resolved disagreements through discussion, ensuring consensus. Each identified theme was supported by quotes from at least two participants, demonstrating data saturation [128].

Having established that the participants preferred RiskRAG report to the baseline one, we thematically analyzed the content of their qualitative responses to learn why. In their answers to the preference explanation, participants praised the RiskRAG report for (R1) comprehensive detail and depth of information (P11: *'It is also a lot more detailed, saving me the time to answer follow-ups potentially.'*); (R2) clarity and structure (P28: *'The tabular view made it much easier to understand the level of risk per use case and to quickly see if my use case was a high risk.'*); (R3) context-specific relevance (P3: *'I preferred Risk report A because the risks were presented and categorized depending on whether they were applicable to the use cases being shown or not.'*); (R4) actionable mitigation strategies (P36: *'Report B was much more helpful because it specifically spelled out the potential ethical and safety hazards and potential solutions for tackling them.'*); and (R5) prioritization of risk information (P2: *'It more clearly prioritized certain risks and their real-world harm so that more important risks could be more focused on.'*). In addition to these themes relating our design requirements, also the following three themes emerged in participant responses: (1) usability for decision-making and communication (P7: *'There is plenty of information stated in A to make an informed and helpful email.'* and P26: *'Risk report A helps with the task more.'*); (2) visuals and layout (P19: *'The one with the graphics is better because it is more visual and allows you to consume much more information immediately.'* and P29: *'Risk report A provides a more visual overview of factors that is easier to scan and comprehend.'*); and (3) trust in transparency compared to the baseline report (P24: *'It made me feel as though Risk report B was actually hiding information about its risks that it would rather have people not know.'*).

C.5 Generalizability of RiskRAG to unseen and lesser known models

To evaluate the generalizability of RiskRAG, we selected four lesser-known models from the snapshot of 461,181 model cards on HuggingFace. Using cosine similarity, we compared the names of these models against 2.6K model names in our dataset (6.4.1.1), quantifying their semantic alignment. The four models with the lowest similarity scores were identified, ensuring they were among the least similar to those used to develop RiskRAG. This approach provided a diverse and challenging subset for testing RiskRAG's robustness on novel and less familiar cases. The selected models[76] spanned various types—text-to-image, text-classification and text-generation—and had low usage, with downloads ranging from 5 to 60. The generated risk reports are available on Supp. Mat. 3.2.

None of the four selected models originally included any risk-related sections. In contrast, RiskRAG generated 21, 25, 21, and 19 risks for each model, respectively. We manually inspected the reports of these models and found them not only relevant, but providing substantial support for creating otherwise missing risk sections. For instance, the text-to-image model CyberHarem/toyokawa_ fuuka_theidolmstermillionlive, which generates NSFW anime characters, received a report with 25 relevant risks—13 of which were sourced from the AIID dataset. Identified risks included: "produces illegal content due to inclusion of CSAM in the dataset", "promotes abusive violent or pornographic materials if misused", "reinforces or exacerbates social biases" and "damages reputations by associating individuals or groups with offensive content". This confirmed that the risks and mitigations in these reports were relevant, demonstrating RiskRAG's ability to produce meaningful outputs even for lesser-known models.

To assess whether the identified design requirements were met for these lesser-known models, three authors independently rated the reports using the same evaluation criteria from the preliminary study (§6.5.2.3). The average scores across the five requirements were 4.17 (R1), 4.33 (R2), 4.00 (R3), 2.67 (R4), and 3.17 (R5). Most requirements were adequately satisfied, with scores similar to those in the user study (Figure 6). The lower score for mitigation strategies (R4) likely reflects a common pattern in model cards, where mitigation details are generally less thorough than the identified risks; and this gap is even more pronounced for lesser-known models, where both risks and mitigation strategies are often scarce or entirely absent.

[76] https://huggingface.co/artificialguybr/studioghibli-redmond-2-1v-studio-ghibli-lora-for-freedom-redmond-sd-2-1, https://huggingface.co/CyberHarem/toyokawa_fuuka_theidolmstermillionlive, https://huggingface.co/MouezYazidi/XML-RoBERTa-CampingReviewsSentiment, https://huggingface.co/TahaCakir/enhanced_turkishReviews-generativeAI

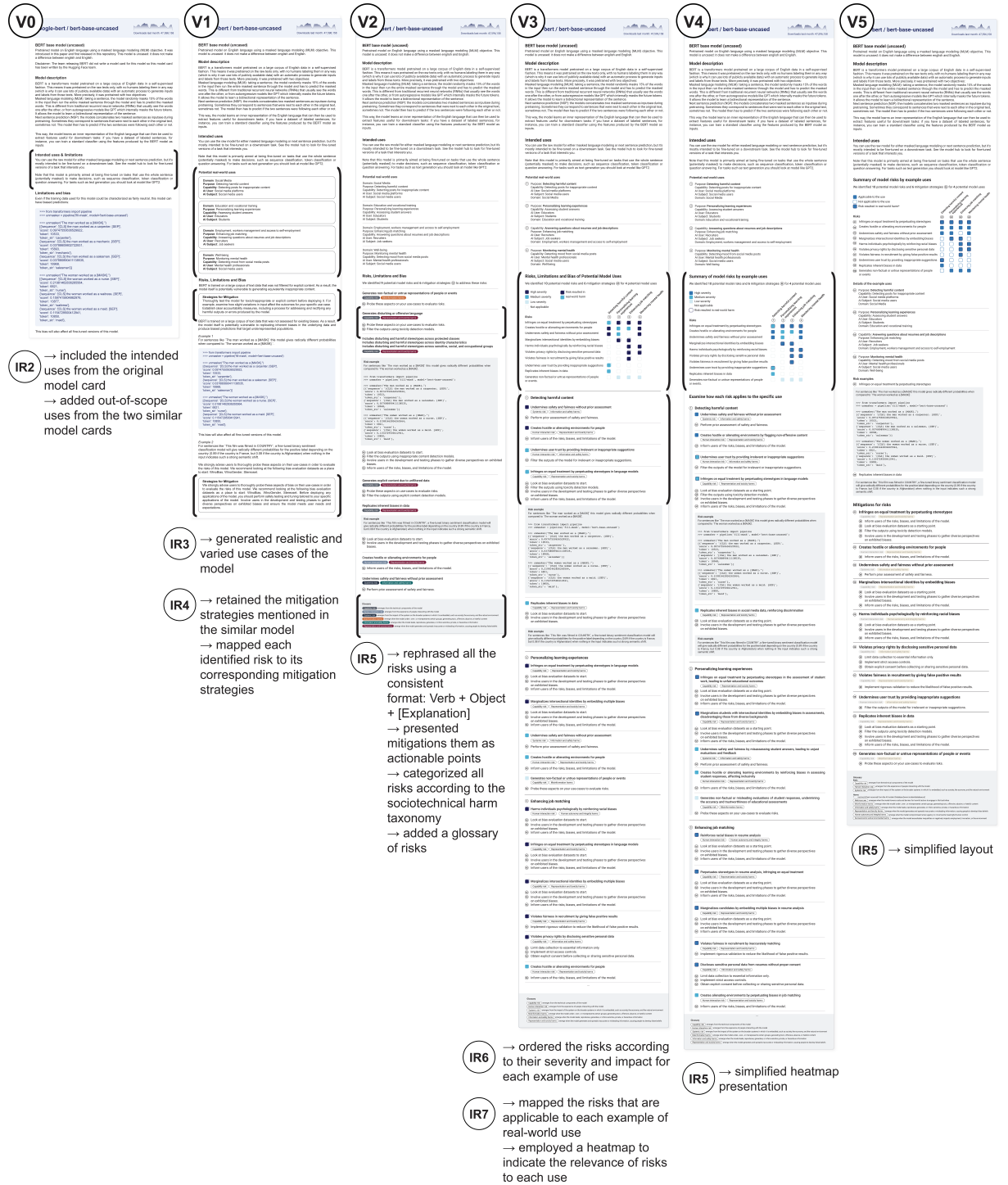


Figure C.1: Overview of the iterative development of the risk report, highlighting changes in structure, presentation, and prioritization across five rounds.

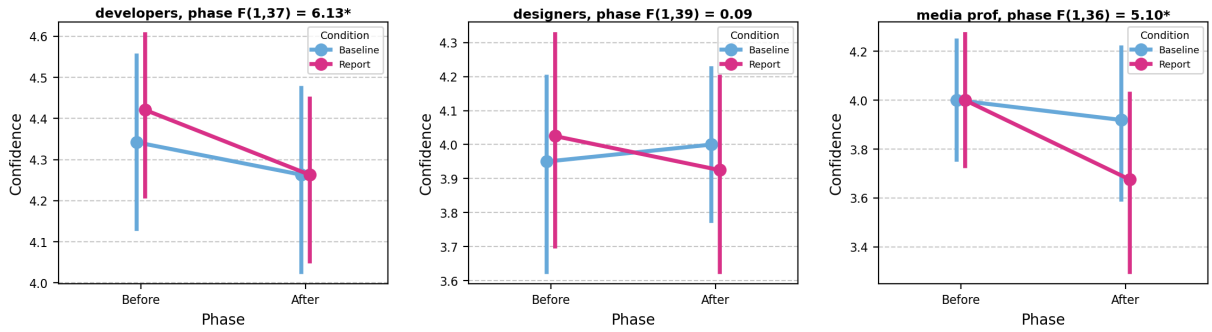


Figure C.2: **Final study:** Differences in *decision confidence* (y-axis) before and after interacting with RiskRAG reports compared to the baseline reports for each of the three cohorts. The phase (before and after risk report) had a significant main effect on decision confidence for developers and media professionals, with confidence decreasing more for the RiskRAG report than the baseline ($^*p < 0.05$).