Selection-Oriented AI: The Role of HCI in Supporting Solutions to Explainability, Plagiarism, and Diversity in Global Scholarship Selection

Neil Natarajan

New College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2024

Acknowledgements

First and foremost, I extend my gratitude to my primary supervisor, Reuben Binns, for his dedication, patience, and wisdom. I would also like to thank my secondary supervisor, Nigel Shadbolt, for his efficient wisdom.

I would like to thank my coauthors and co-conspirators: Sruthi Viswanathan, for teaching me the basics of HCI and for keeping me sane; Thomas Serban Von Davier, for his infinite willingness to read each other's work; Ulrik Lyngs, for teaching me statistics; Elías Hanno, for making much of my action research possible; Kadeem Noray, for an astonishingly deep knowledge of the economics of talent; Logan Gittelson, for writing code that I couldn't have; and Elijah Mayfield, for teaching me how to write a thesis.

I would like to thank my rotating cast of assessors: Max Van Kleek, Jun Zhao, Marina Jirotka, and Tim Miller.

I would like to thank my research community, for fostering an environment of collaboration and support. Thank you Jake Stein, Tyler Reinmund, Lize Alberts, Laura Csuka, Jumana Baghabrah, Sarah Aldaweesh, Sarah Alromaih, Tala Ross, and Helen Gee.

I would like to thank my fiancée for her copyediting prowess; my parents, for their eternal belief in me; and my sister, for teaching me that we go far together.

Abstract

Selecting people for opportunities like jobs, universities, loans, or scholarships pervades and shapes society. And while processes exist for people to make these decisions at scale, these processes are unequipped to handle the elevated demands of modernity. The work in this thesis explores the use of data-driven Decision Support Tools (DSTs) to improve selection processes, focusing on two global scholarship programmes.

We frame our investigation in terms of the *Decision Matrix* framework, categorising decisions by stage (in process or ex post) and stakes (high or low). We then explore using existing AI tools as DSTs, focusing on post-hoc explainable AI and generative AI detectors. We find them ineffective for in-process decisions but useful ex post. We engage in participatory design to create six design prototypes to assist with in-process decision-making, with a focus on diversity. Participants demonstrated enthusiasm for using these tools across the Decision Matrix. To validate this enthusiasm, we implemented one design as a technology probe and evaluated its impact. The selected cohort's diversity and performance improved, demonstrating the tool's ability to support high-stakes in-process decisions.

Our findings highlight the need for data-driven and AI-based DSTs across the Decision Matrix. We propose *Selection-Oriented AI*, a design paradigm focused on the social goals of selection, and provide design recommendations. We conclude with a call for AI-driven DSTs that balance practitioners' needs while optimising selection outcomes for social benefit.

Contents

Glossary

HCI . . . . . Human-Computer Interaction is a subfield of Computer Science that deals primarily with how people interact with computers and to what extent computers are or are not developed for successful interaction with human beings. This thesis is a work of HCI.

PD . . . . . Participatory Design is a paradigm within HCI that engages participants as co-designers in an iterative design process, recognising the user as ideally positioned to understand user needs and preferences. Research outputs are usually designs and design recommendations driven by careful analysis of user feedback. Much of the work in this thesis is inspired by the PD paradigm.

AR . . . . . Action Research is a family of methods within HCI that engages a group of practitioners as co-researchers and co-participants in the research process; in this case, preparation is only one part of the research process, while action and reflection are equally valuable. Research outputs are ordinarily learnings that arise from the action. Much of the work in this thesis is inspired by the AR paradigm.

VSD . . . . . Value-Sensitive Design is a family of methods within HCI engaging participants, where particular values of participants are elicited and used as a guide for design. Research outputs are usually designs and design recommendations driven by careful analysis of user values. This thesis engages with VSD in supporting diversity.

HCC . . . . . Human-Centred Computing is a subfield of Computer Science that designs and develops computer systems around the needs and desires of a group of humans, thus 'centring' that group of humans. This thesis's central contribution (Selection-Oriented AI) is offered in contrast to HCC.

AI . . . . . Artificial Intelligence is variously defined as the study of intelligent behaviour in computers [177], as computational requirements for tasks like perception or reasoning [95], or as large

models such as ChatGPT or DALL-E [63]. AI is often construed as definitionally aspirational, i.e., it is taken as a given that current computer systems are not AI [177]. In this thesis, any computer system that can be said to exhibit behaviour similar to human intelligence is included, and all work herein seeks to build or evaluate AI tools.

XAI . . . . . Explainable Artificial Intelligence is a subfield of AI that develops and assesses explanations that make AI systems more legible to a group of humans. Chapter 4, in particular, engages in a debate over the usefulness of XAI.

GenAI . . . . . Generative Artificial Intelligence is a subfield of AI that develops and assesses AI systems, usually large machine learning models, that generate new data, such as text, images, or audio. Chapter 5 concerns itself with GenAI and the detection of GenAI.

Diversity . . . In its broadest sense, diversity refers to variety, difference, or heterogeneity within a given collection of entities. The seminal definition by Page describes it as: "The heterogeneity of elements in a set about a class that takes different values, such as species in an eco-environment, or ethnicity in a population" [133]. While this definition is broad enough for contexts such as ecology, a more nuanced understanding is required in the context of applicant selection (see Chapter 2.3.1).

HCAI . . . . . Human-Centred Artificial Intelligence is a subfield of HCC that concerns itself with AI systems, rather than all computer systems. This thesis's central contribution (Selection-Oriented AI) is offered in contrast to HCAI.

Selection . . . Occurs in a variety of forms throughout society, from recruitment to matchmaking. In this thesis, selection refers exclusively to the processes of scholarships and other academic or talent investment opportunities, with a primary interest in selection processes for social benefit rather than organisational benefit.

Selector . . . . Practitioners responsible for making selection decisions, both direct and supporting, within selection teams and organisations. These practitioners are referred to as selectors, and tools are built to support their decision-making processes.

SOAI . . . . . Selection-Oriented AI, defined in this thesis, is a family of methodologies designed to achieve the social values of properly

selecting scholars. In contrast to HCAI, which would centre the point of view of various stakeholders, SOAI orients itself around the social benefits of selection, deviating from the point of view of the selectors, applicants, and other stakeholders when their values differ.

1 Introduction

Contents

1.1 Motivation

Consider a prestigious global scholarship program receiving thousands of applications from aspiring scholars worldwide. The selection committee faces a complex decision: from this diverse pool of talented individuals, they must choose a cohort that will best advance the program's mission to develop future leaders who will tackle global challenges. Unlike a simple ranking of academic merit, this selection process involves balancing multiple, sometimes competing values. Should they prioritise applicants with the highest test scores, those who demonstrate the greatest potential for social impact, candidates from underrepresented regions, or individuals who have overcome significant adversity? The committee's approach will depend on

how they interpret the program's goals and what they believe constitutes the "best" cohort, and this decision carries profound implications for both the selected scholars and the communities they will eventually serve.

This "selection problem", choosing whom to include, echoes throughout society. Employers select whom to hire. Creditors select whom they lend to. Universities select whom to admit. Scholarship programmes, like our hypothetical one, select whom to award. An organisation's goals and values determine its selections. An employer may seek the best candidate for a specific task, while a creditor seeks debtors likely to repay loans. Universities and scholarship programmes, however, often differ. Their selection is not for organisational benefit but for a social one [179]. Thus, instead of selecting only applicants who yield the best returns to the organisation, they seek to select those who are most deserving, will learn the most, have the greatest need, whose presence will benefit others, or who will use their education to most improve society.[1]

Although a wealth of research explores when and how algorithms can make hiring or lending decisions (and whether they should) [69, 96, 145, 157, 159], little research explores how algorithms can support selection decisions in the scholarship context. The research that exists in the university context often likens this problem more to hiring than to scholarship [121, 159, 163]. Furthermore, just as research on hiring and lending finds flaws in many applications of algorithms [69, 139, 145], the scant research on human-led scholarship and university selection finds similar problems [159].

In a world flattened by the global proliferation of technology [54], new global scholarship initiatives aim to select scholars from every part of the world. These programmes offer applicants worldwide access to educational resources that may have previously been inaccessible, but they also exacerbate the problems found in related work. If these programmes can select the "best" cohorts, they can deliver on their stated missions to improve the world by broadening access to elite higher

[1] Hirers and lenders are often bound by law to select the most deserving, and they occasionally select those who will benefit others. However, these institutions are generally motivated to maximise profits within legal bounds [157].

education and training scholars to solve the world's most significant problems. This thesis aims to enable that mission by supporting these selection processes with algorithmic decision-support tools.

1.2 Scope of the Thesis

Global scholarship programs, such as the Programme A and Programme B programmes, with whom we conducted this research, face daunting challenges in their mission to select and cultivate future leaders. These challenges include navigating different interpretations of the "best" cohort, ensuring fair comparisons of applicants from diverse global contexts, and mitigating risks such as applicants using generative AI to misrepresent their aptitude. Existing low-tech decision-making systems are often unequipped to handle these newfound complexities [94]. While selection practitioners (selectors) design innovative solutions, they often lack easy access to the information needed to robustly support their decisions.

With now all selection problems being equal, to properly define this thesis's scope, we must distinguish between different types of candidate selection, and narrow our scope to the kind of problem at hand. We identify two primary categories of selection that differ fundamentally in their approach and objectives:

Strict Quota Selection involves assembling groups with specific, predefined compositions to fulfil particular functional requirements. Examples include selecting a sports team where specific positions must be filled (e.g., two defenders, two midfielders, and one goalkeeper) or hiring a team of specific roles (e.g., one software engineer and one salesperson). In these contexts, selectors know in advance how many individuals with particular characteristics they need. The selection process is constrained by these structural requirements, and success requires that the final composition match the predetermined template, but is otherwise evaluated by evaluating each individual independently.

Aspirational Selection aims to meet broader diversity and excellence goals across an entire intake rather than strictly constraining each individual selection decision. This approach characterises most scholarship, university admissions, and similar social programmes. Selectors work toward aspirational targets (e.g., achieving geographic diversity or representing underrepresented groups) but do not have strict quotas. Here, these compositional elements are a goal, rather than a constraint. Success is evaluated by examining whether the cohort reflects the programme's values and mission, which requires considering the group holistically.

This thesis focuses primarily on *Aspirational Selection*, as global scholarship programmes like Programme A and Programme B primarily use *Aspirational Selection* to achieve diversity and excellence. We designed the methods and tools herein to support selection processes where values-based decision-making predominates over structural composition requirements and where the "best" cohort is determined by balancing multiple competing objectives rather than fulfilling predetermined quotas.[2]

Values In scholarship selection, a programme's values are the fundamental principles that guide decisions about which applicants to select. Common values include: *academic excellence* (prioritising applicants with high test scores or grades); *potential for social impact* (favouring candidates with compelling visions for addressing global challenges); *diversity* (seeking applicants from a variety of backgrounds); *need* (supporting those with limited access to education); *integrity* (valuing truth-seeking and truth-telling); or *resilience* (recognising applicants who have overcome adversity). These values often conflict; an applicant with top academic credentials may not show the greatest potential for social impact, and those with the greatest need may not come from the most underrepresented regions. In practice, these conflicts lead to vastly different understandings of the "best"

[2]Historically, some programmes have used *Strict Quota Selection* as a means of operationalising the aspirational goals of the programme in structures that allow stakeholders to evaluate the programme, but this process has fallen out of favour and is not used by either the Programme A or Programme B programmes.

cohort [189]. Selectors must balance these competing values to select cohorts that best embody their programme's mission.

We term the primary decision to overcome this conflict *Selection*: choosing the most apt cohort of applicants according to a specific organisation set of values. This central decision is supported by many subordinate decisions, such as: "What criteria make one applicant (or cohort) more apt than another?" and "How can we apply these criteria to select the most apt cohort?" Each subordinate decision is itself supported by further decisions regarding programme purpose, metrics, and their application.

The scope of this thesis is the development and evaluation of Decision Support Tools (DSTs) to enhance *Selection* processes within global scholarship programmes. Specifically, it focuses on building and assessing DSTs that address three critical challenges selectors face:

1. How to ensure that AI recommendations are interpretable and trustworthy (Chapter 4).

2. How to maintain integrity when applicants may use generative AI (Chapter 5).

3. How to understand, operationalise, and foster diversity within selected cohorts (Chapters 6 and 7).

We conducted this research in collaboration with Programme A and Programme B. While this thesis engages with the ethical dimensions of AI-supported selection, it does not aim to provide definitive solutions to all underlying societal inequities. Instead, it seeks to improve the tools and processes available to selectors. We acknowledge work outside this direct scope, such as critical theory perspectives on selection, in Chapter 8.

1.3 Research Questions

This thesis is guided by the following central research questions:

RQ1: How can Decision Support Tools (DSTs) be effectively designed, implemented, and evaluated to aid selectors in global talent investment programs in making values-driven selection decisions that balance multifaceted objectives like excellence, diversity, and fairness, across both aspirational quota and cohort-based selection paradigms?

RQ2: In what ways do emerging AI technologies, particularly explainable AI (XAI) and generative AI (GenAI), impact the decision-making processes, integrity, and perceived legitimacy of scholarship selection, and what frameworks or interventions can mitigate potential harms while leveraging benefits for selectors?

RQ3: How can complex and evolving conceptualizations of diversity be translated into practical support mechanisms within DSTs to assist scholarship programs in achieving their diversity-related goals, and what is the real-world efficacy of such mechanisms when deployed in live selection processes?

1.4 Contributions

The contributions of this thesis are twofold. There are meta-level conceptual distinctions introduced, and also some substantive contributions associated with body chapters.

The meta-level contributions of this thesis are:

- A list of decision points facing scholarship programmes uncovered through longitudinal HCI research with Programme A and Programme B.
- The SOAI paradigm for designing AI systems that support one of the social benefits of good selection processes.

- A set of design recommendations for designers seeking to apply SOAI to build a DST to support selectors.

However, this thesis is composed of several papers that make more specific, core contributions to support tools seeking to solve issues of Explainability, Plagiarism, and Diversity. These are detailed in the relevant chapters but are also described here.

Explainability

- Quantitative findings indicating that the problem of explanation-induced unwarranted trust extends to generic post-hoc justifications, but that such criticism only applies in process (Chapter 4).
- Qualitative findings that post-hoc explanations, properly presented, can make useful ex-post DSTs (Chapter 4).

Plagiarism

- An evaluation of GenAI detectors GPTZero and Originality.ai on Programme A's Cycles X and Y application data (Chapter 5).
- The Decision Matrix framework for evaluating the suitability of AI systems as support tools for differing decision points.
- A case study using GPTZero to support two decision points facing Programme A (Chapter 5).

Diversity

- The Diversity Triangle, categorising diversity-related themes according to our three definitions of diversity uncovered through inductive thematic analysis (Chapter 6).
- Six design prototypes developed through PD for supporting the diversity needs of a given organisation (Chapter 6).

- Design recommendations grounded in PD for system implementers supporting the diversity needs of a given organisation (Chapter 6).

- A field deployment of Prototype 6.3c to the Programme A selection process selecting several hundred finalists from a pool of several thousand demonstrating the efficacy of this prototype in practice (Chapter 7).

- A demonstration of a hypothetical application of Prototype 6.3c as an ex-post DST (Chapter 7).

1.5 Thesis Structure

Chapter 2 serves as an extended introduction and background chapter, including situating this thesis in related work. Following this, Chapter 3 explores the paradigms that guide research design throughout this thesis, lists methods used throughout the thesis, and ties these methods to specific chapters.

Chapter 4 responds to common criticisms of post-hoc XAI and explores this approach as a scholarship selection DST via PD workshops with selectors from Programme A. Chapter 5 engages selectors from Programme A in an AR process and explores the role of generative AI in selection decisions. Chapter 6 engages selectors from both Programme A and Programme B in participatory design to explore selector notions of diversity and potential ways to support these considerations; this chapter ultimately develops 6 design prototypes. Chapter 7 implements one of these prototypes in a field deployment with Programme A, evaluates that deployment, and explores other applications of the technology.

Chapter 8 discusses the scope of the thesis, including references to critical theory work falling outside the scope; the SOAI paradigm; design recommendations that developers can use to follow SOAI; methodological and technical limitations of the work herein; and this thesis's broader significance in a quickly changing landscape.

1.6 Papers

1.6.1 Archival and Under Review

- Neil Natarajan, Sruthi Viswanathan, Reuben Binns, Nigel Shadbolt. 2024. " 'Diversity is Having the Diversity': Unpacking and Designing for Diversity in Applicant Selection". Under review at CHI 2025.
- Neil Natarajan, Reuben Binns, Ulrik Lyngs, Nigel Shadbolt. 2024. "XAI: Misleading In Process, but Useful Post Hoc". Under review at CHI 2025.
- Neil Natarajan, Elías Hanno, Logan Gittelson, Reuben Binns, Nigel Shadbolt. 2024. "What Are Generative AI Detectors Good For? Evaluating and Implementing with the Decision Matrix". Under review at CHI 2025.

1.6.2 Peer Reviewed

- Neil Natarajan et al. "Detecting Generative AI Usage in Application Essays". en. In: *Generative AI and HCI workshop at CHI 2024*. 2024. URL: https://generativeaiandhci.github.io/papers/2024/genaichi2024_9.pdf
- Neil Natarajan. "Human-AI Collaboration in Recruitment and Selection". In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. Ed. by Edith Elkind. Doctoral Consortium. International Joint Conferences on Artificial Intelligence Organisation, Aug. 2023, pp. 7089–7090. URL: https://doi.org/10.24963/ijcai.2023/819
- Neil Natarajan et al. "Trust Explanations to Do What They Say". en. In: *Human-Centered AI Workshop at NeurIPS 2022*. 2022. URL: https://openreview.net/pdf?id=mzsPCefDaY5

2
Background and Context

Contents

2.1 The Challenge of Global Scholarship Selection

This thesis addresses fundamental challenges in designing and implementing algorithmic decision support tools (DSTs) for global scholarship selection. As these programmes have grown, traditional processes have become increasingly difficult to manage, sparking interest in algorithmic solutions to enhance efficiency and fairness [94]. Scholarship programmes aim to provide long-term societal benefits by selecting the "best" applicants, but what "best" means depends on the organisation's theory of change—whether it prioritises the future contributions of scholars, the

inherent social good of providing opportunity, or the productivity gains from diverse perspectives [39, 75, 128].

However, applying algorithmic DSTs to this context is uniquely challenging. Unlike in hiring or university admissions, global scholarship selection must contend with extraordinary diversity in applicants' backgrounds, educational systems, and cultural contexts [179]. While algorithms have much potential to support unique challenges in this new context, attempting to introduce algorithms into the selection process creates significant technical and ethical hurdles. Critics raise valid concerns about algorithmic bias and the dehumanisation of the selection process [18, 46], though human-led processes are equally vulnerable to critique [2].

This complex landscape raises several key research questions:

1. How can we design DSTs to support fair and effective selection across diverse global populations?

2. How can we provide meaningful explanations that enhance, rather than undermine, human decision-making?

3. How should we adapt to emerging technologies like generative AI?

4. How do we balance competing values of merit, diversity, and social impact?

These questions are fundamentally socio-technical, requiring an interdisciplinary approach that draws on computer science, social science, and critical technology studies.

2.2 Historical and Societal Foundations

Contemporary challenges in algorithmic selection must be situated within a broader historical context of discrimination, civil rights, and the evolving role of technology in society. We undertake this situation here.

2.2.1 The Pursuit of Equity in Selection

The modern emphasis on diversity and fairness is rooted in decades of social and political struggle. The U.S. Civil Rights Movement of the 1950s and 1960s challenged systemic exclusion, leading to affirmative action policies designed to redress historical inequities [5, 119]. This movement, initially focused on the under-representation of women and racial minorities in the U.S., has evolved into a global concept of diversity that encompasses a wide variety of identities [126].

Global scholarship programmes inherit this legacy. Contemporary programmes like Programme A and Programme B, with whom this thesis engages, grapple with similar tensions as they seek to operationalise diversity without perpetuating tokenism. This history provides critical context for understanding why modern selectors prioritise both individual merit and diversity within and across cohorts.

2.2.2 Technology, Power, and Social Justice

The pursuit of equity is further complicated by the role of technology. As scholars like Benjamin [15] and Noble [127] have shown, technological systems can encode and perpetuate societal biases, reinforcing existing power structures. Algorithms, while promising objectivity, often reflect the historical inequities present in their training data and can enact "categorical violence" by forcing fluid identities into rigid demographic bins [21, 156].

Thus, technology can democratise access to resources, but it can also entrench new forms of exclusion [48, 149]. Building technology that serves the former requires attention to what Jasanoff [74] calls the 'co-production' of science, technology, and social order: the ways in which technological systems both shape and are shaped by social values and political structures. The development of DSTs is therefore not a purely technical exercise but also a political process demanding careful attention to power, representation, and accountability.

2.3 Some Conceptual Pillars of Selection-Oriented AI

In designing DSTs for selection via a process we term Selection-Oriented AI, we tackle several conceptual pillars. We do not claim our list of pillars to be complete, but contend here that it is sound; that is, when designing DSTs for selection processes of the kind explored in this thesis, designers should pay attention to the following.

2.3.1 Diversity: From Theory to Practice

Diversity is a central, yet ambiguous, concept in selection. It draws from multiple intellectual traditions: social psychology distinguishes between demographic and cognitive diversity [133]; organisational behaviour links diversity to team performance [134]; and political philosophy values diversity for a flourishing democracy [109, 185]. Economically, diversity is often quantified via entropy metrics and linked to productivity [128].

These varied motivations lead to a nebulous definition. This thesis starts with Page [133]'s general definition: "The heterogeneity of elements in a set", but acknowledges its limitations when applied to people. In practice, diversity considerations often involve nuanced concepts like **group diversity** (equitable representation), **intersectional diversity** (recognising overlapping identities), and **cognitive diversity** (differences in thinking) [38, 68, 116]. As we explore in Chapter 6, selectors often simultaneously navigate these different meanings.

2.3.2 Fairness: Competing Ideals in Algorithmic Systems

The algorithmic fairness literature provides a critical lens for evaluating selection processes. A key tension exists between **individual fairness** (treating similar applicants similarly) and **group fairness** (achieving parity across demographic groups) [46, 79]. This manifests in debates over distributive versus procedural justice. While distinct from diversity, fairness is often intertwined in practice; strategies to enhance one can impact the other, creating complex trade-offs that selection systems must navigate [187].

2.3.3 Explainable AI (XAI): Promise and Peril

As machine learning models have grown more complex, XAI has emerged to make their decisions more transparent and trustworthy. This quest for interpretability is not new; it dates back to early rule-based expert systems [162]. Modern XAI distinguishes between **intrinsically interpretable models** (like decision trees) and **post-hoc explanations** (most often applied to "black-box" models that lack native interfaces for interpretability) [117, 154].

However, post-hoc explanations come with the risk of **automation bias**, where users over-rely on automated advice, and the "trust paradox", where plausible but misleading explanations can increase unwarranted trust [90, 120]. As we investigate in Chapter 4, this raises critical questions about when and how explanations should be used in high-stakes selection decisions.

2.3.4 Generative AI: A New Frontier of Challenge

The recent rise of powerful generative AI, particularly Large Language Models (LLMs) like GPT-4, has introduced a new layer of complexity to selection [130, 173]. Students are increasingly using these tools to write application essays, challenging traditional notions of authorship and assessment [41]. While institutions have been quick to issue policies, these often lack clarity and are difficult to enforce, as current detection tools have significant limitations [115, 140]. This new reality forces a re-evaluation of the role of written assessments in selection, a central theme we explore in Chapter 5.

2.4 Situating the Thesis

This thesis is positioned at the intersection of HCI, AI, and social science, engaging with a sparse body of literature on AI in global scholarship selection. While much has been written on AI in hiring or university admissions, the unique challenges of global scholarships—particularly the need to compare applicants across vastly different contexts—remain under-explored. Our work builds on and extends research

in algorithmic fairness, explainable AI, and the economics of talent selection [82, 98], applying these lenses to a novel empirical setting.

To ground this research, we engaged in longitudinal partnerships with two global scholarship programmes: **Programme A** and **Programme B**. Both are innovative programmes seeking to leverage AI in their selection processes. Through Action Research, Value-Sensitive Design, and Participatory Design, we worked with selectors from these organisations to identify key challenges and co-design solutions. These collaborations surfaced several families of subordinate decisions that form the empirical core of this thesis: challenges related to explainability (Chapter 4), applicant use of generative AI (Chapter 5), and cohort diversity (Chapters 6 and 7). A full list of the decision points addressed can be found in Table 2.1.

Table 2.1: This table enumerates relevant challenges facing selectors from the Programme A and Programme B selection teams. Challenges are drawn from discussions with selectors, where descriptions are framed in terms of decisions these programs make.

Challenges	Chapter(s)	Description	Supporting Information
Refinement	4 and 7	A programme may refine its scoring algorithm each year to better score applicants.	Explanations of perplexing AI-generated scores; information about implications of scoring methods for cohort diversity
Diligence	5	A programme may make holistic decisions about when and how to consider applicants.	Information about which essays (and which parts of essays) were written by GenAI; information about whether the GenAI-written passages are hallucinations.
Partners	5	A programme may determine whether to continue referral relationships, which encourage and support applicants.	Whether any referral organisations' affiliated applicants use GenAI disproportionately.
Pipeline	5	A programme may decide whether to modify their application material or process.	Information about the usage of GenAI throughout the application pipeline.
Gameability	5	A programme may decide how to modify their application material or process.	Information about how AI-generated essays are scored under the current application process.
Disqualification	5	A programme may decide to disqualify an applicant that violates their application guidelines.	Information about whether essays violate application guidelines around GenAI usage.
Diversity	6 and 7	A programme may make cohort-level decisions regarding the diversity of their cohort.	Information about the diversity of possible cohorts.
Contribution	6 and 7	A programme may make decisions about which applicants to move forward based on their contribution to diversity.	Information about the impact of including different applicants on cohort diversity.

3 Methodology

Contents

3.1 Methodological Framework of the Thesis

This thesis employs a mixed-methods approach that combines computational evaluation with qualitative investigation of social and organisational contexts. This section outlines the key methodological frameworks that inform the research, grounding it in established practices while tailoring them to the unique demands of studying AI in global scholarship selection.

3.1.1 Core Research Traditions

The methodological framework of this thesis is grounded in several key traditions from **Human-Computer Interaction (HCI)** and related fields. Our approach is rooted in **Participatory Design (PD)**, which centres the experiences and needs of users (in our case, we primarily consider scholarship selectors and their organisations) [20, 83]. This tradition recognises that effective Decision Support Tools (DSTs) cannot be designed in isolation but must emerge through collaborative engagement with stakeholders [23]. To facilitate this, we also incorporate **Action Research (AR)**, which emphasises conducting "research with, rather than on, people" [22, 65]. This approach is particularly vital in scholarship selection, where organisational dynamics and constraints heavily shape technological interventions [102].

However, engaging only the stakeholders in these decisions risks satisfying individuals while failing to achieve the broader social aims of selection. Thus, guiding these participatory approaches is **Value Sensitive Design (VSD)**, which ensures that we design algorithmic systems with explicit attention to human values and their implications [172]. This is crucial in selection contexts where decisions have profound impacts on individuals' lives.

Finally, though not central to our approach, we adopt a **critical perspective on algorithmic systems**, following scholars like Noble [127] and Roy [153] in exploring the risks of implementing algorithmic systems. This perspective requires examining not only technical performance but also the social and political implications of DSTs, helping to reveal how they may reproduce or challenge existing inequalities.

Integrating Methodological Traditions While each of these traditions offers a unique lens, the strength of this thesis's methodology lies in their integration. We use Value-Sensitive Design (VSD) as our foundational framework to conceptually identify and prioritise the human values at stake in scholarship selection, such as fairness, diversity, and transparency. This values-centred groundwork then informs our practical engagement. We employ Participatory Design (PD) and Action Research (AR) as engines for translating these values into tangible interventions.

Through PD, we collaborate directly with selectors to co-create tools and frameworks grounded in their lived experiences and organisational realities. AR provides the iterative structure for deploying, testing, and refining these interventions in a real-world setting, ensuring our research is not only theoretically sound but also practically relevant and responsive to the evolving needs of our partner organisations. This synergistic approach allows us to move from abstract principles (VSD) to collaborative creation (PD) and finally to real-world impact (AR).

3.1.2 Evaluation Methods

To assess both technical performance and socio-technical implications, this thesis employs a multi-faceted evaluation strategy.

For **computational evaluation**, we use standard machine learning metrics, including accuracy, precision, recall, and F1-scores, to assess algorithmic performance. Recognising the limitations of these metrics in capturing fairness concerns, we also employ fairness-aware evaluation methods, including stratification along demographic lines and calibration analysis [107].

For the **human-centred evaluation of XAI**, we follow the framework of Doshi-Velez and Kim [44]. This involves both functional evaluation (measuring how explanations affect task performance) and human-grounded evaluation (measuring how explanations align with human processes). This dual approach recognises that technical performance alone is insufficient for understanding the social implications of algorithmic systems.

3.1.3 Critical and Reflexive Stance

Drawing on critical algorithmic studies, the thesis adopts a reflexive approach that examines not only the technical properties of algorithmic systems but their social, political, and ethical implications [160]. This includes attention to questions of power, representation, and accountability that are often overlooked in purely technical approaches to algorithm design. These approaches emphasize collaboration

and shared ownership of the research process, recognizing that effective technological solutions must emerge from genuine engagement with user communities.

3.2 Specific Methods Used

3.2.1 Online Surveys

The practice of running online surveys to gather quantitative data is well-established and often used both within and without HCI [9, 18, 47, 87, 104, 137, 142, 187]. Chapter 4 makes use of one such survey. We use Prolific Academic to gather participants and Formr to administer our survey [7, 18]. We follow Caldwell et al. [30] in designing our survey based on a power analysis of the statistical tests we intend to run on the output data.

3.2.2 Design Workshops

Chapters 4 and 6 both make use of group design workshops to refine and evaluate design prototypes. Both follow an experience-prototype methodology [27], and incorporate a few specific methodologies.

Both chapters follow Zimmerman and Forlizzi [190]'s scenario speed dating approach, which sees participants rapidly applying different design prototypes to (real or hypothetical) scenarios.

Gatian [56] has researchers asking participants to choose a favourite among a series of options as a means of comparison, while Griffiths et al. [61] brings this method to HCI. Chapter 6 makes use of this method.

3.2.3 Individual Interviews

Chapter 6 makes use of one-on-one interviews with participants to first elucidate participant understanding of diversity. In these interviews, we incorporate several methods.

Knapp et al. [83]'s 'crazy 8s' exercise sees participants give eight feature requests in eight minutes. Ordinarily, this exercise is done with a writing surface, but we have participants do this verbally.

Blythe [20] introduces the concept of design fiction, where participants more detail their ideal app. We adapt this to create a "magic app", capable of doing anything the participant desires and asking the participant to describe this app.

3.2.4 Quantitative Analysis

Chapters 4, 5, and 7 rely on several standard statistical tests. Primarily, we use Student's t-test [114], the Analysis of Variance (ANOVA) [114], Pearson's test of correlation [158], Tukey's Honestly Significant Difference test [80], and the Receiver Operating Characteristic curve [62]. Additionally, we develop a permutation test in Chapter 7 based on Good [58].

3.2.5 Qualitative Analysis

Chapters 4 and 6 engage in inductive thematic analyses of their qualitative results. In doing so, we follow the methodology introduced by Braun and Clarke [23] and developed in Braun and Clarke [24, 25] and *Thematic Analysis / SAGE Publications Ltd* [169].

3.3 Research Design

Chapters 4, 5, 6, and 7 all detail studies conducted according to different research paradigms and employing different methodologies. Each chapter contains a self-encapsulated section on research design. However, Table 3.1 provides a high-level overview of the methods and paradigms employed in each chapter.

	Chapter 4	**Chapter 5**	**Chapter 6**	**Chapter 7**
Participatory Design	Yes		Yes	
Action Research		Yes		Yes
Value-Sensitive Design			Yes	
Online Surveys	Yes			
Design Workshops	Yes		Yes	
Individual Interviews			Yes	
Quantitative Analysis	Yes	Yes		Yes
Qualitative Analysis	Yes	Yes	Yes	

Table 3.1: This table indicates which methods and paradigms are employed in each core research chapter.

4

XAI: Misleading In Process, but Useful Ex Post[3]

Contents

[3]This chapter is based on a paper written in concert with Reuben Binns, Ulrik Lyngs, and Nigel Shadbolt. The paper is currently under review as: Neil Natarajan, Reuben Binns, Ulrik Lyngs, and Nigel Shadbolt. 2024. "XAI: Misleading In Process, but Useful Ex Post." Under review at CHI 2025.

4.1 Motivation

When exploring decision support tools for global scholarship selection, Explainable AI (XAI) is a natural starting point. Scholarship selection often involves complex algorithmic scoring across multiple dimensions, from cognitive assessments to interview performance. Proponents offer XAI tools as a way to support a decision-subject's right to an explanation, aiming to empower individuals and improve decision-making in sensitive fields [59]. While a wealth of research explores XAI in these contexts [11, 117, 176], others caution against the blind application of these tools [14, 88]. This chapter explores the potential of post-hoc interpretability to support selection-related decisions and considers how these tools might be applied in different contexts.

4.2 Introduction

For a comprehensive background on XAI, automation bias, and the history of interpretable AI systems, see Chapter 2.3.3. This section focuses on the specific critiques relevant to our research questions.

Despite its promise, post-hoc XAI has faced increasing criticism as a decision support tool. Many studies centre on how these systems affect user trust and decision-making, revealing several distinct concerns. We will examine these critiques in detail.

Lipton [100] offers a foundational critique, arguing that well-intentioned explanation design may yield "misleading but plausible" explanations. Their work suggests that providing an explanation can create a false sense of understanding, leading users to trust the system's outputs even when that trust is unwarranted. This critique is particularly relevant in selection contexts, where the stakes of misplaced trust are high.

Building on this, Miller [111] identifies a more fundamental issue: post-hoc explanations often serve to justify the underlying AI models and their outputs rather than enabling users to make their own informed decisions. They argue

that these explanations can create a form of "algorithmic authority," where the explanation itself becomes a tool for legitimising the model's decisions, potentially undermining human agency.

The evidence for these concerns is mixed but troubling. Lai and Tan [90] found that explanations can increase user trust in AI systems, but this trust may not be well-calibrated to the system's actual performance. Similarly, Jacobs et al. [73] demonstrated that explanations can lead to over-reliance on AI, particularly in high-stakes contexts. These findings suggest that the problem of misplaced trust operates through multiple mechanisms, from creating a false sense of understanding to establishing algorithmic authority.

In response to these critiques, the field has seen a significant shift away from post-hoc approaches. Kumar et al. [88] argues that post-hoc explanations often fail to provide meaningful insights into model behaviour, while Bastounis et al. [14] demonstrates that these explanations can be manipulated to justify incorrect decisions. This has led to the development of new paradigms, such as Miller [111]'s evaluative AI, which focuses on helping users evaluate model outputs rather than explaining them, and Karimi et al. [78]'s causal models, which aim to provide more actionable insights.

But have we been too hasty in rejecting post-hoc methods? Underlying this shift is the assumption that such approaches will be deployed to increase trust in particular outputs, even when that trust is unwarranted. But is this always the context of their use? Could trust-inducing explanations be deployed for other purposes, such as for evaluating models and decision-making processes after the fact?

To analyse the benefits and risks of post-hoc interpretability tools, we develop a decision-stage distinction (also explored in Chapter 5). That is, rather than partitioning examples by type of explanation, we partition by the timing of its use. While the former is a property of the algorithms themselves, the latter is a property specific to a given use case [124]. A 'post-hoc' explanation is one generated after a model has been trained, in contrast to an ante-hoc explanation or inherently interpretable model. This is distinct from when the explanation is presented to a

user. We distinguish between two timings: **in process**, where AI outputs and post-hoc explanations support a human decision-maker concurrently with the primary decision, and **ex post**, where the primary decision has already been made and XAI informs second-order decisions about the process itself. For example, selectors might use ex post explanations between application cycles to refine their procedures for the next cycle [98]. This distinction helps us frame Chapter 4's two research questions:

(RQ1) Do post-hoc explanations, when used as in-process DSTs, induce unwarranted trust in users?

(RQ2) If post-hoc explanations induce unwarranted trust in process, could they still be useful ex post?

To answer RQ1, we run an online study to discern whether the problem of unwarranted trust is specific to certain types of post-hoc explanations. We investigate two popular methods, SHapley-based Additive exPlanations (SHAP) [103] and Scoped Rules (Anchor) [152], to see if they induce unwarranted trust. We also investigate a 'Confidence' explanation consisting of the model's confidence statistic to determine if the problem applies more generally to any information that could increase positive perceptions of the AI's performance. We ask participants to *estimate a person's salary* [84] or *predict whether someone will be severely delinquent in making a credit payment* [37] with the help of an AI. We find that SHAP explanations increase unwarranted trust in AI outputs, but so do Confidence explanations. We find no such effect for Anchor. This suggests the problem of unwarranted trust is not unique to a specific kind of XAI and is rather a symptom of generic post-hoc justifications.

Having identified a core problem with some kinds of in-process XAI, we use participatory design workshops to consider whether they might have redeeming features if deployed at the ex-post stage. We focus on SHAP, as the critiques from Lipton [100] and Miller [111] are most germane to it. We contend that, in an ex-post context, inducing unwarranted trust is less problematic, as primary decisions have already been made. We ask participants to *refine a scholarship selection algorithm*

with the help of SHAP-based explanations. Through these workshops, we find that while SHAP may induce unwarranted trust in specific model outputs, it can still be useful to drive process change in organisations.

Our primary contributions are:

1. Quantitative findings indicating that the problem of explanation-induced unwarranted trust extends to generic post-hoc justifications, but that such criticism only applies when explanations are used in process.

2. Qualitative findings that post-hoc explanations, when properly presented, can be useful ex-post DSTs.

4.3 Experimental Study (In Process)

4.3.1 Research Questions

Our online study seeks to answer RQ1:

(RQ1) Does post-hoc XAI used as an in-process DST induce unwarranted trust in users?

To do this, we compare three alternate conditions: Lundberg and Lee [103]'s SHAP explanations, Ribeiro et al. [152]'s Anchor-based explanations, and a Confidence condition consisting of the model's intrinsic confidence measurement. We measure trust in the AI system in two ways: attitudinal trust, measured by self-report, and behavioural trust, measured by the participant's decision to follow the AI system's recommendation. We also measure the change in trust before and after the explanation and compare this change across the three conditions. These measurements are done across two tasks. Each participant sees six cases, with the explanatory and task conditions held constant.

4.3.2 Methodology

4.3.2.1 Participants

Participants were recruited via Prolific Academic's standard sampling method restricted to the United States.[4] They were paid at a rate of $15 per hour. Participants were first shown an information sheet detailing the study's methodology and what was being asked of them. They were then asked to give informed consent. After consenting to participate in the study, participants were routed to Formr, our chosen survey design and hosting platform, to complete the online study.[5] All data collected was anonymous and was stored on secure servers. Ethics review was performed by the University of Oxford's Central University Research Ethics Committee.

4.3.2.2 Tasks

We chose tasks that are familiar to laypeople and have a well-defined but difficult-to-ascertain ground truth from a gamut of well-known algorithmic decision-making tasks as two particularly related to *Selection* [94, 138, 144]: *estimating a hypothetical person's salary* based on census information, and *predicting whether someone will be severely delinquent in making a credit payment*. These tasks mirror key aspects of scholarship selection:

1. The salary estimation task parallels the evaluation of an applicant's socioeconomic status and need, a common consideration in scholarship selection [179]. Just as selectors must assess financial need, this task requires evaluating complex socioeconomic factors to make a determination.

2. The credit delinquency task reflects the challenge of predicting an applicant's likelihood of 'success' (according to program-defined criteria) [159]. Selectors

[4] `www.prolific.co`
[5] `www.formr.org`

must often assess future potential or the likelihood that a candidate will give back to the community.[6]

We use two datasets: the Adult dataset from the 1994 US Census for the former task and the Give Me Some Credit dataset for the latter [37, 84]. These datasets contain a mix of germane and demographic attributes, including sensitive ones. In both tasks, the participant aims to accurately make a determination with the help of the AI system and one of several possible explanations of its output.

These tasks engage participants in making decisions based on potentially biasing information, as there are no direct causal links establishing these attributes as definitive predictors, but some exhibit germane correlations with outcomes, while others do not. This complexity mirrors real selection processes, where selectors must distinguish germane information from noise (often including sensitive or demographic attributes). These tasks therefore serve as a realistic sandbox to gauge the baseline decision-making of human raters and the impact of XAI on that process.

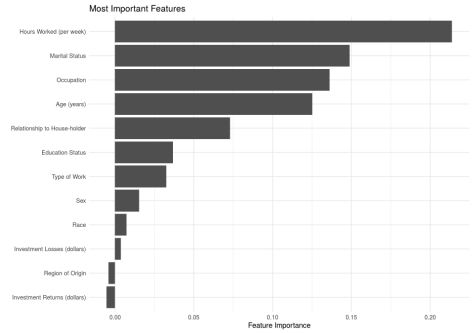
Our aim is not to endorse the fairness of these predictive tasks or the models themselves, but to investigate how XAI methods influence user trust and decision-making in such ethically complex environments. The handling of these ethical issues was a key consideration in the study design. The study protocol, including the use of these tasks and datasets, was reviewed and approved by the University of Oxford's Central University Research Ethics Committee.

In our analyses, we index these tasks as *Salary* and *Credit*. Each participant only receives one task to complete across all six cases.

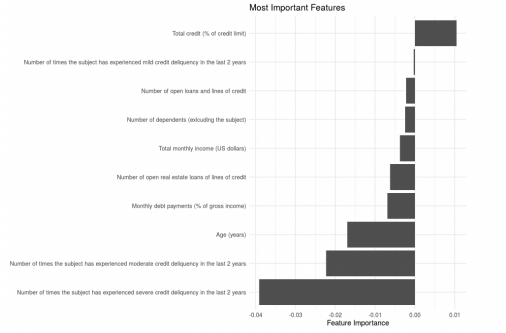
4.3.2.3 Models

In both tasks, we construct a predictor model using random forests and augment this predictor with three different explanatory conditions. Our random forest classifier achieves 86% test accuracy on the Adult dataset and 93% test accuracy on the Give Me Some Credit dataset. We use a SHAP explainer to produce one of our

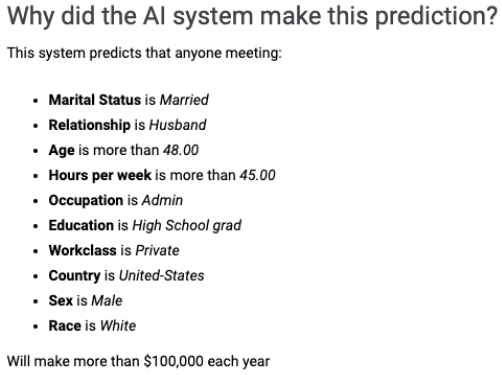
[6]Educational institutions often receive funding from former students and thus have a financial interest in ensuring that recipients are likely to give back [164]. While this is ordinarily not an explicit factor in scholarship selection, it may nevertheless play a role.



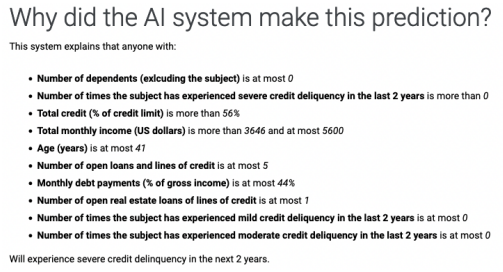
(a) SHAP explanations for *Salary*



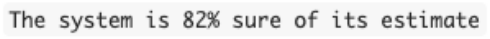
(b) SHAP explanations for *Credit*



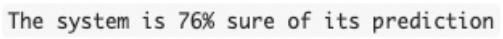
(c) Anchor explanations for *Salary*



(d) Anchor explanations for *credit*



(e) Confidence explanations for *Salary*



(f) Confidence explanations for *Credit*

Figure 4.1: This figure shows sample explanations for all cases. Larger images and more detailed descriptions of explanations can be seen in Appendix D.1.

explanatory conditions, and an Anchor explainer to produce another; our final explanatory condition is an intrinsic explanation produced by the random forest model. Figure 4.1 shows sample explanations produced by these methods. These ultimately form three explanatory conditions: SHAP, Anchor, and Confidence. Note that each participant only receives one model of explanation throughout all 6 cases.

4.3.2.4 Design

Both tasks rely on the same 3-between-by-2-within design using repeated measures to capture the same data before and after the presentation of each explanation. The between-subjects factor determines which model is used to generate the explanation a given participant will receive. The within-subjects factor is the repeated-measures

'explanation presence' factor. This is either 'before explanation' or 'after explanation', indexed $before$ or $after$. A flowchart of the study design can be found in Figure 4.2.

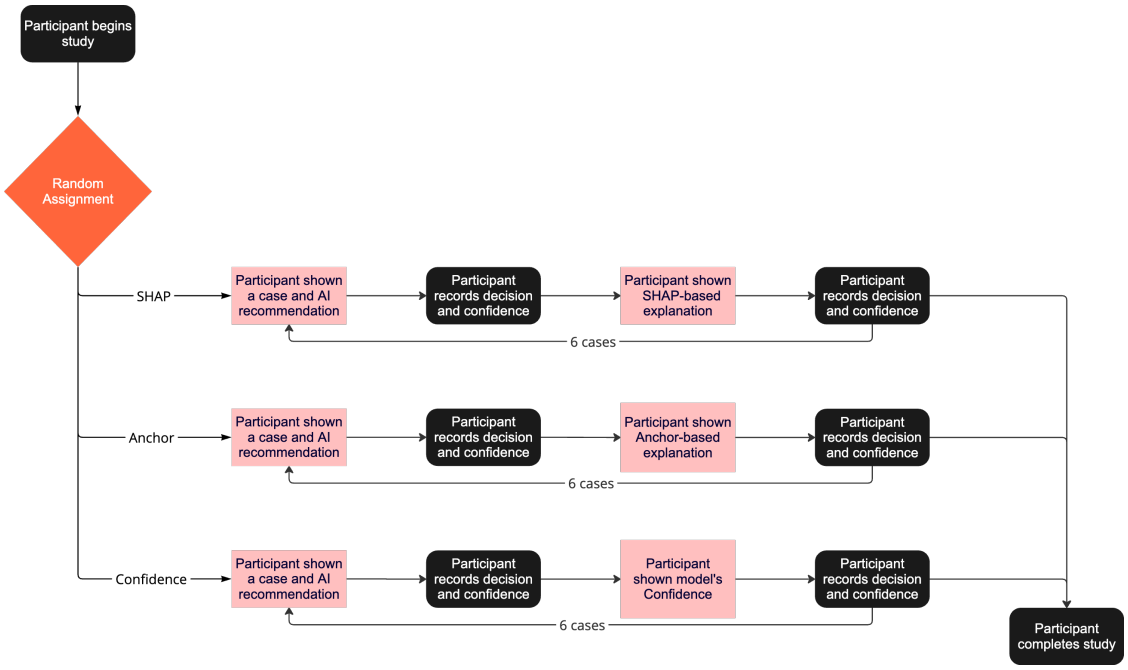


Figure 4.2: Participants in the online study are sorted into six buckets, where each bucket is segregated by explanatory condition and task and shown a brief description of the task (i.e., each participant sees only one of the explanations in Figure 4.1). Then, each participant is shown 6 cases. In each case, participants are shown an applicant profile and an AI output. Participants are asked to agree or disagree with the AI output. Then, participants are given explanations based on their explanatory condition scores. They are then asked again to agree or disagree with the AI output.

4.3.2.5 Questions and Variables

Each participant was shown a brief explanation of the task in question and was then asked to complete six cases, with participants given a random mix of correct and incorrect AI outputs. In each case, participants are first shown a table identifying the subject of the case and the AI's binary determination. They are then asked to make their own determination for the case and rate their confidence in that choice. They also rate their trust in the AI's output. These ratings are on sliding scales (discretised to 20 points).

We code the participant's determination as a binary variable, y_{human}:

$$y_{human} := \begin{cases} \text{This person makes more than \$100,000 per year} & (Salary) \\ \text{This person will experience severe credit delinquency} & (Credit) \end{cases} \tag{4.1}$$

The two sliding scale responses are coded as $selfconfidence$ and $trust_{attitudinal}$ and have values between 1 and 20. These are defined as:

$$confidence := \begin{cases} \text{How confident are you in your estimation?} & (Salary) \\ \text{How confident are you in your prediction?} & (Credit) \end{cases} \tag{4.2}$$

$$trust_{attitudinal} := \begin{cases} \text{How much do you trust the AI's estimation?} & (Salary) \\ \text{How much do you trust the AI's prediction?} & (Credit) \end{cases} \tag{4.3}$$

As we ask all questions in both the $before$ and $after$ conditions, we collect six responses from each participant in each case: y^{before}_{human}, $selfconfidence^{before}$, $trust^{before}_{attitudinal}$, y^{after}_{human}, $selfconfidence^{before}$, and $trust^{after}_{attitudinal}$. We additionally have the binary variables y_{True} and y_{AI} that are the true value and the AI output of the dependent variable.

In addition to these, we define $agreement^x$ as:

$$agreement^x := \begin{cases} 1 & \text{if } y^x_{human} = y_{AI} \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

and $correct_x$ as:

$$correct_x := y_x = y_{True} \tag{4.5}$$

We define $trust^{before}_{behavioural}$ and $trust^{after}_{behavioural}$ to be the extent to which the participant's confidence agrees with the AI output:

$$trust^x_{behavioural} := \begin{cases} confidence^x & \text{if } agreement^x \\ 1 - confidence^x & \text{otherwise} \end{cases} \tag{4.6}$$

Finally, to reason about the change in a variable due to the explanation, we define 'Δ' constructs for all variables with a $before$ and an $after$ as:

$$\Delta variable := variable^{after} - variable^{before} \tag{4.7}$$

so, e.g.:

$$\Delta trust_{attitudinal} := trust^{after}_{attitudinal} - trust^{before}_{attitudinal} \tag{4.8}$$

4.3.2.6 Data Analysis

We preregistered many of our analyses. Though we also include some post-hoc analysis below, we wish to delineate between the two types of analyses. The former are listed in full here.

We first wish to test for the presence of unwarranted trust. To do this, we measure the difference between ratings in each condition in the cases where the AI output is incorrect, i.e. where $\neg correct_{AI}$. To do this, we run three one-sided t-tests [30] to determine:

$$\Delta trust \geq 0 | \neg correct_{AI} \tag{4.9}$$

for both $\Delta trust_{behavioural}$ and $\Delta trust_{attitudinal}$. Note that this is identical to repeated-measures t-tests on the $trust_{behavioural}$ and $trust_{attitudinal}$ variables where $\neg correct_{AI}$. Following this, as this is also a between-subjects experiment, we wish to compare the varying effects of different explanation methods in this case. Thus, we run two between-subjects ANOVAs [30] on $\Delta trust_{behavioural}$ and $\Delta trust_{attitudinal}$ across the explanatory conditions again filtered on $\neg correct_{AI}$. When these ANOVAs have significant results ($p < 0.05$), we run Tukey's Honestly Significant Difference (HSD) test [30].

Finally, in our Salary Estimation survey, we observed a strong positive correlation between our two trust variables, though we did not preregister it for that study. Indeed, correlation analysis is a well-documented method for confirming that two measurements indeed measure the same concept [118, 181]. Thus, we additionally included the calculation of Pearson's correlation between $trust_{behavioural}$

and $trust_{attitudinal}$ and between $\Delta trust_{attitudinal}$ and $\Delta trust_{behavioural}$ in our preregistration for the Credit Delinquency Prediction survey.

Power Analysis We run power analyses using Caldwell et al. [30]'s Superpower. Specifically, we desire sufficiently powerful results in our primary analysis. We select a moderate effect size of interest (Cohen's f) of 0.15 (yielding group means -0.15 and 0.15 with unit variance), and target a power of at least 0.90. We note that using a one-way t-test at $p = 0.05$ and assuming a sample size of 200, we get power far above 0.90. We also test for the ANOVA assuming unit variance and group means of -0.15, 0.15, and 0.15, respectively. Under these conditions, we achieve a power of 0.90 with 200 samples per condition. We ask each participant a total of 6 questions, and the AI output is incorrect in slightly less than half of them. To achieve 200 samples per condition, therefore, we aim to recruit a total of roughly 66 participants per condition, or roughly 200 participants in total.

Preregistration We have preregistered analyses for both of our tasks in the OSF registries [122].

4.3.3 Results

In both tasks, though we originally set 200 as our target participants, some participants did not complete our task following Prolific Academic's guidelines. Data from these participants were marked incomplete and removed from consideration.

After this removal, we had a total of 192 participants complete the Salary Estimation study. These were split randomly into our three explanatory groups. By gender, 115 were Male, 76 were Female, and 1 did not provide gender information. By ethnicity, 137 were white, 10 did not provide ethnicity, and the remaining 45 were split among non-white ethnicities. Our participants were an average of 36.7 years old, with the youngest being 18 and the oldest 74. Each applicant completed an introductory page and six cases. The average completion time for these tasks was 7 minutes 43 seconds, the minimum was 2 minutes 25, and the maximum was 36 minutes 46.

We had a total of 197 participants complete the Credit Delinquency Prediction study. These were similarly split into groups. By gender, 106 were Male, 90 were Female, and 1 did not provide gender information. By ethnicity, 143 were white, 11 did not provide ethnicity, and the remaining 43 were split among non-white ethnicities. Our participants were an average of 38.4 years old, with the youngest being 20 and the oldest 77. Each applicant completed an introductory page and six cases. The average completion time for these tasks was 7 minutes 53 seconds, the minimum was 2 minutes 17, and the maximum was 30 minutes 13.

SHAP and Confidence Increase Unwarranted Trust We first run the one-sided t-tests on the two trust variables (*attitudinal* and *behavioural*). I.e., we test:

$$\Delta trust_x > 0 | \neg correct_{AI} \text{ for } x \in \{behavioural, attitudinal\} \tag{4.10}$$

for both *Salary* and *Credit* across SHAP, Anchor, and Confidence. A positive F statistic here indicates $trust^{after} > trust^{before}$ and a negative F statistic indicates $trust^{after} < trust^{before}$, but, as these are one-sided tests, p-values will only be meaningful when $F > 0$. This test was preregistered in both of our tasks [122]. Table 4.1 contains the results of these analyses.

Table 4.1: These one-sided t-tests test for $\Delta trust > 0 | \neg correct_{AI}$ for all explanatory conditions and both tasks. We find that SHAP and Confidence increase unwarranted trust in the AI system.

Task	Explanation	Variable	Test Statistic	p Value
Salary	Anchor	$\Delta trust_{behavioural}$	0.509	0.306
		$\Delta trust_{attitudinal}$	0.165	0.434
	SHAP	$\Delta trust_{behavioural}$	**3.811**	**< 0.001**
		$\Delta trust_{attitudinal}$	−0.886	0.812
	Confidence	$\Delta trust_{behavioural}$	**2.196**	**0.015**
		$\Delta trust_{attitudinal}$	0.945	0.173
Credit	Anchor	$\Delta trust_{behavioural}$	1.396	0.082
		$\Delta trust_{attitudinal}$	−2.364	0.990
	SHAP	$\Delta trust_{behavioural}$	1.516	0.066
		$\Delta trust_{attitudinal}$	**2.475**	**0.007**
	Confidence	$\Delta trust_{behavioural}$	**1.835**	**0.034**
		$\Delta trust_{attitudinal}$	0.940	0.174

This indicates that SHAP and Confidence appear to lead users to trust the AI system more when that system is wrong. We find this result more strongly for behavioural trust than attitudinal trust in all but one test.

Notably, Anchor does not follow this pattern and instead shows no significant increase in either attitudinal or behavioural trust on these one-sided t-tests. (However, as we explore in Section 4.3.3, they may show a significant decrease.)

Different Explanation Styles Have Different Effects on Unwarranted Trust We have shown already that SHAP and Confidence induce unwarranted trust relative to no explanation; we now show that there is a significant difference in the effect of some explanatory conditions relative to others. To do this, we examine the $\Delta trust_{behavioural}$ and $\Delta trust_{attitudinal}$ variables across explanatory conditions with an ANOVA test. For this test, we filter on $\neg correct_{AI}$:

$$\begin{aligned} &\Delta \text{ any}(trust_{x1,x2} \neq trust_{x1,x3}) | \neg correct_{AI} \\ &\quad \text{for } x1 \in \{behavioural, attitudinal\} \\ &\quad \text{and } x2, x3 \in \{SHAP, Anchor, Confidence\} \end{aligned} \tag{4.11}$$

This test was preregistered in both of our tasks [122]. Table 4.2 contains the results of these analyses.

Table 4.2: These ANOVAs compare $\Delta trust$ between SHAP, Confidence, and Anchor to indicate where significant differences exist. We find two statistically significant differences, indicating that we should focus post-hoc analyses on these two.

Task	Variable	Test Statistic	p Value
Salary	$\Delta trust_{behavioural}$	**3.671**	**0.026**
	$\Delta trust_{attitudinal}$	0.925	0.397
Credit	$\Delta trust_{behavioural}$	0.066	0.936
	$\Delta trust_{attitudinal}$	**6.213**	**0.002**

Note from Table 4.2 that in the Salary Estimation task, we find no significant results for our ANOVA $trust_{attitudinal}$, but do find significant results for $trust_{behavioural}$. However, in the Credit Delinquency Prediction task, we find significant results

for our ANOVA $trust_{attitudinal}$, but none for $trust_{behavioural}$. We examine these two findings separately.

SHAP Increases Behavioural Trust More than Anchor in the Salary Estimation Task We now show that:

$$\Delta trust_{behavioural,SHAP,salary} > \Delta trust_{behavioural,Anchor,salary} | \neg correct_{AI} \quad (4.12)$$

Note that, while the result of the ANOVA test in ctab:delta-trust-anova supports that there are indeed statistically significant differences in the three group means of the $\Delta trust_{behavioural}$ variable, it does not specify which means are greater and which are less. For an indication of which means are greater, following our preregistered protocol for significant ANOVA results, we turn to Tukey's Honestly Significant Difference (HSD) test as a post-hoc test in Table 4.3. As we found significant results in the ANOVA test of $\Delta trust_{behavioural}$ filtered on $\neg correct_{AI}$, we restrict our post-hoc analysis to this variable.

Table 4.3: Tukey's HSD test compares $\Delta trust_{behavioural,x}$ in *Salary* with $\neg correct_{AI}$. We find that SHAP increases behavioural trust in incorrect AI outputs more than Anchor.

Explanation A	Explanation B	Variable	Test Statistic	p Value
SHAP	Anchor	$\Delta trust_{behavioural}$	**2.310**	**0.022**
Confidence	Anchor	$\Delta trust_{behavioural}$	0.855	0.599
SHAP	Confidence	$\Delta trust_{behavioural}$	1.455	0.198

As can be seen in Table 4.3, we observe a significant difference in the mean of $\Delta trust_{behavioural}$ between the SHAP and Anchor conditions with $\neg correct_{AI}$, but we do not observe a significant difference between the other conditions. This indicates that, beyond increasing behavioural trust in incorrect AI outputs, SHAP increases behavioural trust in incorrect AI outputs *more* than Anchor.

Anchor Decreases Unwarranted Attitudinal Trust Relative to SHAP and Confidence in the Credit Delinquency Prediction Task Note that we found a large negative F for $\Delta trust_{attitudinal}$ in the Anchor case in the Credit Delinquency Prediction portion of Table 4.1 – an effect that is not significant due to the one-sidedness of our tests. However, as we found only positive F values for $\Delta trust_{attitudinal}$ in the SHAP and Confidence cases, we might expect that Anchor has a negative effect on $\Delta trust_{attitudinal}$ relative to SHAP and Confidence. Indeed, the result of the ANOVA test in Table 4.2 supports that there are indeed statistically significant differences in the three group means of the $\Delta trust_{attitudinal}$ variable in this task, though it again does not specify which groups are different, or how.

For an indication of which means are greater, following our preregistered protocol for significant ANOVA results, we turn again to Tukey's HSD test as a post-hoc test in Table 4.4. As we found significant results in the ANOVA test of $\Delta trust_{attitudinal}$ filtered on $\neg correct_{AI}$, we restrict our post-hoc analysis to this variable. We test:

$$\begin{aligned} &\Delta trust_{behavioural,x1,credit} > \Delta trust_{behavioural,x2,credit} | \neg correct_{AI} \\ &\quad \text{for } x1, x2 \in \{SHAP, Anchor, Confidence\} \end{aligned} \tag{4.13}$$

$$\begin{aligned} &\Delta \text{ any}(trust_{x1,x2} \neq trust_{x1,x3}) | \neg correct_{AI} \\ &\quad \text{for } x1 \in \{behavioural, attitudinal\} \\ &\quad \text{and } x2, x3 \in \{SHAP, Anchor, Confidence\} \end{aligned} \tag{4.14}$$

Table 4.4: Tukey's HSD test compares $\Delta trust_{attitudinal}$ across explanations in *Credit* with $\neg correct_{AI}$. We find that Anchor decreases unwarranted trust relative to SHAP and Confidence.

Explanation A	Explanation B	Variable	Test Statistic	p Value
SHAP	Anchor	$\Delta trust_{attitudinal}$	**1.213**	$<$ **0.001**
Confidence	Anchor	$\Delta trust_{attitudinal}$	**1.030**	$<$ **0.001**
SHAP	Confidence	$\Delta trust_{attitudinal}$	0.183	0.708

As can be seen in Table 4.4, we observe a significant difference in the mean of $\Delta trust_{behavioural}$ between the Anchor condition and both other conditions, but

we do not observe a significant difference between SHAP and Confidence. This indicates that, relative to both other conditions, Anchor actually *reduces* attitudinal trust in the AI output.

Note that this does not prove that Anchor reduces attitudinal trust relative to no explanation. For this analysis, we will need another t-test. As we did not preregister this test, an analysis of this phenomenon is included in the exploratory proportion of our results below.

Behavioural and Attitudinal Trust are Highly Correlated It should be noted that some patterns observed for $trust_{behavioural}$ do not hold for $trust_{attitudinal}$ and vice-versa. However, while they are mathematically distinct constructs, they are both intended to measure the same underlying phenomenon. We apply Pearson's correlation analysis across all explanatory conditions in both the before- and after-cases. We also perform this analysis on $\Delta trust_{attitudinal}$ and $\Delta trust_{behavioural}$.

For this analysis, we do not filter out positive cases. Rather, we consider all cases together. Results can be seen in Table 4.5.[7]

Table 4.5: Pearson's test shows a high correlation between $trust_{attitudinal}$ and $trust_{behavioural}$ in both tasks. This correlation extends to the relationship between $\Delta trust_{attitudinal}$ and $\Delta trust_{behavioural}$.

Task	Variable A	Variable B	Test Statistic	p Value
Salary	$trust_{attitudinal}$	$trust_{behavioural}$	**0.630**	$<$ **0.001**
	$\Delta trust_{attitudinal}$	$\Delta trust_{behavioural}$	**0.265**	$<$ **0.001**
Credit	$trust_{attitudinal}$	$trust_{behavioural}$	**0.612**	$<$ **0.001**
	$\Delta trust_{attitudinal}$	$\Delta trust_{behavioural}$	**0.179**	$<$ **0.001**

Note that, though the attitudinal and behavioural trust variables display different behaviours in other analyses, Table 4.5 indicates that they are indeed highly correlated across both of our tasks. Furthermore, though the correlation between the $\Delta trust$ is more modest, it is still statistically significant. These together leave little doubt that $trust_{attitudinal}$ and $trust_{behavioural}$ measure related, and perhaps

[7]This analysis was only partially preregistered; we did not register this analysis in the Salary Estimation task, but we did in the Credit Delinquency Prediction task [122].

even identical, concepts. In other words, when a participant says they trust the AI output, they generally act accordingly.

Anchor Decreases Attitudinal Trust in AI Outputs Having found no significant result indicating the presence of the hypothesised effect of Anchor explanations on unwarranted trust, we explore what effect Anchor explanations have on end users' trust in incorrect AI outputs. For this analysis, we filter on $\neg correct_{AI}$.[8]

We noted already that SHAP and Confidence appear to increase trust in cases where the AI output is incorrect. However, we noticed no such result for Anchor. However, we did observe an apparent negative effect on $\Delta trust_{attitudinal}$ in the t-tests for the Credit Delinquency Prediction task, though, as the tests were one-sided, we could not confirm the significance. Thus, we repeat this test as a two-sided test, as shown in Table 4.6. We also repeat other t-tests on Anchor as two-sided, though, as they all have positive F-values, note that none can yield significant results.

Table 4.6: Two-sided t-tests compare $\Delta trust_{x,Anchor} \neq 0 for x \in \{attitudinal, behavioural\}$ in both tasks. We find that Anchor explanations decrease attitudinal trust in incorrect AI outputs in *Credit*, but all other results are inconclusive.

Task	Explanation	Variable	Test Statistic	p Value
Salary	Anchor	$trust_{behavioural}$	0.509	0.611
		$trust_{attitudinal}$	0.165	0.869
Credit	Anchor	$trust_{behavioural}$	1.396	0.164
		$trust_{attitudinal}$	**−2.364**	**0.019**

Note here that, on the two-sided t-test, we *do* find that the provision of Anchor explanations decreases participant attitudinal trust in incorrect AI output, at least in the Credit Delinquency Prediction task. However, we do not see a similar effect on behavioural trust.

Anchor and SHAP Increase Participant Self-Confidence in their Determinations Noting that behavioural trust is an index variable constructed from $selfconfidence$, we ask: does providing an Anchor explanation increase participant

[8]This analysis was not preregistered.

confidence in their own decisions when the AI output is incorrect? Similarly, we ask this question for both the SHAP and Confidence conditions, following exactly the format of the preregistered one-sided t-tests in Table 4.1, but applied to the variable $\Delta selfconfidence$. Again, we filter on $\neg correct_{AI}$.:

$$\Delta selfconfidence > 0 | \neg correct_{AI} \tag{4.15}$$

Analysis can be seen in Table 4.7.[9] Note for clarity that $selfconfidence$ is the variable indicating participant confidence in their own decisions, and Confidence is the condition in which the explanation consists of the AI's own confidence in its suggestion.

Table 4.7: One-sided t-tests determine whether $\Delta selfconfidence > 0$ where $\neg correct_{AI}$ for all explanatory conditions and both tasks. We find that Anchor and SHAP increase participant self-confidence in their determinations, while Confidence yields inconclusive results.

Task	Explanation	Variable	Test Statistic	p
Salary	Anchor	$selfconfidence$	**2.171**	**0.016**
	SHAP	$selfconfidence$	**1.694**	**0.046**
	Confidence	$selfconfidence$	1.047	0.296
Credit	Anchor	$selfconfidence$	**1.742**	**0.042**
	SHAP	$selfconfidence$	**3.473**	$<$ **0.001**
	Confidence	$selfconfidence$	0.752	0.226

We note that, while Confidence shows no significant effects on either task, participants shown an Anchor or SHAP explanation grow significantly more confident in their prediction, indicating that providing an Anchor or SHAP explanation serves to increase a participant's confidence in their own estimate.

Explanations Impact Trust Differently When the AI Output is Correct Note that ideally calibrated trust would involve both distrusting the AI output when it is wrong and trusting it when it is right. To assess the latter, we now turn to an evaluation of what happens in the cases where the AI is correct, i.e.

[9] This analysis was not preregistered.

$correct_{AI}$. Namely, we conduct two-sided t-tests on both trust variables in all three cases. Table 4.8 contains the results of these analyses.[10]

Table 4.8: These two-sided t-tests compare $\Delta trust$ when $correct_{AI}$. We find that Confidence and SHAP increase trust in the AI system when it is correct, while Anchor decreases attitudinal trust in the AI system, but increases behavioural trust.

Task	Explanation	Variable	Test Statistic	p Value
Salary	Anchor	$trust_{behavioural}$	0.502	0.616
		$trust_{attitudinal}$	**−2.337**	**0.020**
	SHAP	$trust_{behavioural}$	0.295	0.768
		$trust_{attitudinal}$	−1.385	0.168
	Confidence	$trust_{behavioural}$	**2.410**	**0.017**
		$trust_{attitudinal}$	**3.254**	**0.001**
Credit	Anchor	$trust_{behavioural}$	**3.013**	**0.003**
		$trust_{attitudinal}$	**−2.487**	**0.014**
	SHAP	$trust_{behavioural}$	0.207	0.836
		$trust_{attitudinal}$	**3.538**	**0.001**
	Confidence	$trust_{behavioural}$	**2.863**	**0.005**
		$trust_{attitudinal}$	**2.461**	**0.015**

Confidence Explanations Increase Warranted Trust When the AI Output is Correct Note that, for both trust variables and both tasks, the Confidence condition boasts a significant positive $\Delta trust$. In other words, when the AI output is correct, providing confidence in its own prediction increases both behavioural and attitudinal trust towards the AI.

Anchor Explanations Decrease Warranted Attitudinal Trust but Increase Warranted Behavioural Trust When the AI Output is Correct In the Anchor case, it is clear that providing Anchor explanations yields a large decrease in $trust_{attitudinal}$. This, along with the finding that Anchor explanations decrease $trust_{attitudinal}$ when $\neg correct_{AI}$, would indicate that Anchor explanations have an overall negative impact on $trust_{attitudinal}$, regardless of case. Despite this, providing Anchor explanations yields an increase in $trust_{behavioural}$ (though this is only significant in the *Credit* case). This suggests that, though participants report

[10]These analyses were not preregistered.

lower trust in AI outputs when shown an Anchor explanation, they behave as though their trust in the outputs is appropriately calibrated.[11]

SHAP Explanations Increase Warranted Attitudinal Trust in the Credit Delinquency Prediction Task When the AI Output is Correct In the SHAP case, we find a large significant positive F for $trust_{attitudinal}$ when $correct_{AI}$ in the credit delinquency prediction task. However, not only is the effect not mirrored in either test of $trust_{behavioural}$, but the same test on the salary estimation task has a negative F statistic.

4.3.4 Study Findings

Our results indicate that both SHAP and Confidence induce unwarranted trust in the explainee. I.e., on the *Salary* and *Credit* tasks, neither SHAP nor Confidence serves to correctly calibrate trust in AI outputs. Rather, they blindly increase trust in these outputs, encouraging users to incorrectly agree with the AI outputs. While this confirms the cautionary critique of Lipton [100] as applied to SHAP in our domain, our results relating to Confidence suggest this critique is too narrow. Namely, issues of unwarranted trust do not seem confined to what is commonly considered post-hoc XAI but are rather a function of providing any post-hoc justification of the model's output. This suggests that even un-optimised notions of interpretability induce unwarranted trust when provided as justifications of model outputs.

We note that our study into these two explanations yields little insight into *why* participants trust the AI outputs. Indeed, it is unclear from our findings whether this trust arises because the explanations lend an air of authority or complexity, whether they might induce confirmation bias by highlighting features that seem to support the AI's conclusion (a particular concern for feature-attribution methods like SHAP), or whether participants simply lack a strong impetus to question the AI when an explanation is provided. Our study into the Anchor condition may help to shed light on this question, as we find no similar effect for Anchor.

[11] Though $trust_{attitudinal}$ and $trust_{behavioural}$ are closely correlated overall, this represents one instance in which they appear to reveal differences in participant behaviour.

Instead, we find a significant decrease in participants' stated confidence in AI and a simultaneous increase in participant self-confidence in their own decisions. Miller [110] identifies several features that social sciences would suggest make a good explanation. Among them, Anchor explanations are contrastive, counterfactual, and selective, while the SHAP and Confidence conditions lack these properties. It may be that Anchor explanations serve to highlight strange model behaviour that correctly undermines explainee confidence in model outputs.

In short, the problem seems not to be the use of explanations as justification tools, but rather the use of "bad" explanations as justification tools. This raises another question: if SHAP and Confidence are "bad" explanations as in-process justification tools, are they "good" explanations for something else?

4.4 Participatory Design (Ex Post)

4.4.1 Motivation

In Section 4.3's investigation of post-hoc explainable AI, we found that SHAP-based explanations can lead to unwarranted trust when used to justify decisions. It is clear, at least, that we should not use SHAP-based explanations as a DST in this context. However, we do not suggest that SHAP should be discarded entirely. In fact, though *Salary* and *Credit* both prove unsuitable tasks for applications of SHAP-based explanations as DSTs, we suggest that we might still make use of the explanations. In particular, in use cases where explainee trust in the underlying model is not at issue, SHAP's induction of unwarranted trust need not undermine its utility.

It is clear that when making a 'primary' decision (i.e., the same decision the model output seeks to make), human reviewers working from AI outputs are preeminently concerned with whether to agree with or overrule the model's output. However, after making this decision, they are no longer necessarily concerned with whether to agree or disagree with the model's output. Consider the task of *refining a scholarship selection algorithm*, the *Refinement* task from Chapter 2. Scholarship and talent investment programmes ordinarily select cohorts in a series of application cycles [98]. Between application cycles, they seek to examine their previous selection decisions,

and possibly modify their processes to improve these decisions in the future [98]; if AI algorithms are used to support these decisions, the review will naturally include refinements to the AI algorithms. In this case, though, we are no longer concerned with whether the AI was correct; rather, we are concerned with whether the way the AI informs decision-making is conducive to ideal selection pipelines.

We conducted a human-centric study using SHAP explanations as an ex-post explanation tool to help selectors from Programme A (see Appendix A) with the *Refinement* task. Through this participatory design study, we assess SHAP's usefulness based on the insights these explanations provide.

This study aims to answer RQ2:

(RQ2) If post-hoc XAI methods induce unwarranted trust in process, could they still be useful ex post?[12]

In doing so, we restrict our attention specifically to SHAP, as we have already demonstrated its induction of unwarranted trust.

4.4.2 Methodology

4.4.2.1 Programme A's Selection Process

We use the Programme A selection process as a case study for our participatory design study. Programme A is a scholarship and talent investment programme that selects cohorts of approximately N winners and 5N finalists from their pool of applicants. We focus primarily on this final selection of winners from finalists. At this stage, the programme has already undergone a stage of selection in which applicants submit video essays describing a project they completed and complete a cognitive assessment. These video essays are then assessed twice: once by a randomly selected group of the applicant's peers (other applicant), and another time by a group of external experts. These assessments are used once more in the final selection of winners.

[12]Recall the distinction between ex post and post hoc. We use the term 'post hoc' to refer to explanation algorithms that are applied after a model; 'ex post', in contrast, refers to the explanation tools applied in decision support scenarios where the primary decisions have already been made.

Additionally, all finalists are asked to participate in a day-long series of remote workshops where they are asked to complete a variety of tasks in front of a panel of external experts (this panel contains a subset of the external experts who assessed the applicants' video essays). The scores from these workshops are used alongside older metrics to produce algorithmically-generated scores and demographic information is used to test the bias of these scores. All scores, as well as a variety of qualitative factors, are then used to select the final cohort of winners.[13]

4.4.2.2 Our Study

We test SHAP's usefulness in a human-centred context. To do so, we ran a study with scholarship and talent investment selectors (N=8) from Programme A. Though Programme A already used algorithmic scoring to support their decision-making, no selectors possessed experience with post-hoc XAI before the study. The study consisted of two participatory design workshops with said selectors – we term these 'G1' and 'G2'.[14]

Before this study, we obtained informed consent from all participants. As we ran group workshops (and as we do not attribute comments to individual participants), participants were informed that they would not be able to recuse themselves after the study. Participants also gave consent to be recorded, and to have these recordings stored on a secure server. All recording, transcribing, and data analysis was conducted on secure servers. Ethics review was performed by the University of Oxford's Central University Research Ethics Committee.

Both workshops followed an identical protocol. The flow of these workshops is shown in Figure 4.3, and more detail on the protocol followed can be found in Appendix B.1.

In each workshop, participants discussed several cases, each examining a (possibly successful) applicant from a past application cycle who was flagged by programme

[13]The programme has requested we not disclose the details of their selection process, but many particulars of the program are excluded by design.

[14]As Programme A only employs a small number of easily identifiable selection selectors, to preserve the anonymity of the participants, we do not number or identify participants. Rather, we attribute quotes based on the workshop group.

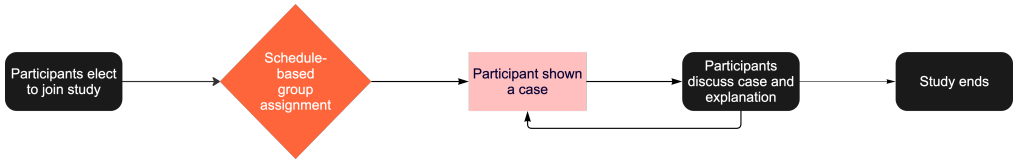


Figure 4.3: Each workshop consisted of a series of cases relating to a past application decision that was flagged by programme reviewers. In each case, participants were shown slides like in Figure 4.4 and were asked to analyse the algorithm itself and whether the case warrants changes to the algorithm in future years.

reviewers for having perplexing algorithm scores relative to other known information. In each case, participants are asked to use visual, SHAP-based explanations to first understand why programme reviewers found these cases worth noting, then to explore why programme reviewers gave the feedback they did and what caused the algorithm's perplexing outputs, and finally to opine on whether the case suggests that changes should be made to the algorithm (or to the selection process as a whole) for future years. The cases themselves have been redacted, as they contain sensitive information about programme applicants, but a sample case can be seen in Figure 4.4.

In analysing our data, we follow Braun and Clarke [23]'s methodology for reflexive thematic analysis.

4.4.3 Results

Our case study yielded two key themes. Firstly, SHAP Explanations yield useful ex-post insights about feature importance. Second, even though SHAP yields useful information, the accessibility of such information depends on careful presentation. We now cover these in depth.

Ex-Post Insights on Important Features In both groups, several useful insights emerged due to the SHAP visualisations. For example, the relationships between scores and contextual factors (e.g., markers of applicants' socioeconomic status) revealed that context plays little to no role in scoring; despite this, an

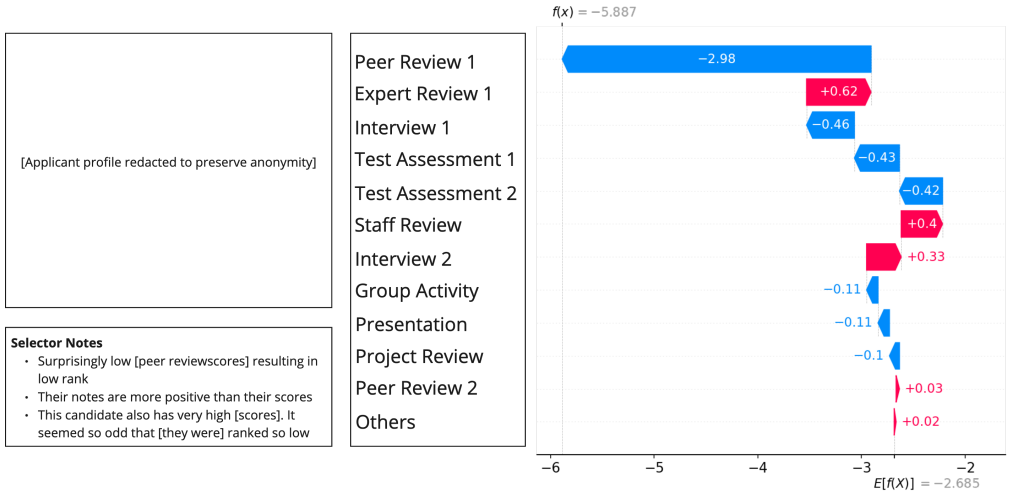


Figure 4.4: Each case explores one applicant from past years chosen by past programme reviewers after being flagged as having perplexing algorithm scores. Each case contains the applicant's profile (overall algorithm scores alongside demographic information; the profile is redacted to preserve applicant anonymity), the programme reviewers' comments, and the SHAP-based explanation (the score names are replaced with generic labels to preserve programme anonymity).

applicant's context has a strong impact on how selectors read scores.[15] E.g., an applicant with high test scores from a poor region of Kenya is more impressive than one with high test scores from a rich part of the United Kingdom. When discussing one applicant who was selected, but had particularly low algorithmic scores, one participant said: "This is one of the candidates that... [was from a] different country and [had] very low income" (G1).

It was also remarked upon that scores calculated based on the assessment of external experts often disagreed with scores calculated based on the assessment of programme applicants. It was discovered here that, contrary to programme expectations, the programme's expert reviews appeared less prone to biases than the peer ones. For one applicant: "There was a question about why his peer and expert review were so different...I think confirms that it's not actually that they were seeing dramatically different things...his peers were dinging him for not seeming like he needed the award" (G1). For another: "I think this applicant has been

[15]Meanwhile, we find in Chapter 6 that selectors consider contextualising applications a key motivation behind considerations of diversity.

significantly brought down by peer reviews; [their] scores are substantially lower than those that were, perhaps, given to [another applicant]" (G1).

Similarly, it was observed that, unlike project reviews, group activities, and test results, the best candidates do not appear to have particularly good interview scores: "I'm seeing also quite a few top-ranked candidates whose interview score was really low" (G1). In some cases, this appeared to create a discrepancy between algorithmically generated overall scores and the participants' perceptions of the best candidates: "[The applicant's] staff reviews imply that [they] should be top 30, and even if you factor in low interview scores...[they] are still pretty low on the algorithm score" (G2).

Presentation is Key Besides insights about the selection process, the workshops yielded direct feedback on how the presentation of SHAP explanations should be improved. One major point was that contextual factors describing an applicant were missing. One participant said, of the explanation: "It doesn't give me the context" (G1). Another from the same group said: "But I think without the context, it's really hard to decipher what's going on here" (G1). This could be interpreted as participants asking for supporting information. However, when the researchers read out the information in the explanation's caption, this cleared up the participant's confusion: "Yeah, that makes sense" (G1). This suggests that, rather than needing more information, participants needed the information presented differently.

Several times, participants were unclear on the meaning of different aspects of explanations: "Should I be alarmed and I see it going blue in the context of this? It's really hard for me to if you threw this at me...to compare" (G1). Participants asked for the more complicated information as a "Pre-reading" (G2), and asked for simpler information, i.e., "Maybe just colour coding things that are positive in one colour and then things that are negative in the other colours" (G2).

One request that several participants echoed was that axes be kept constant, even between different types of scores: "It's also different scaling. So, that massive bar...does not mean the same thing as the massive bar meant last time" (G2).

Another solution suggested to a similar problem was the provision of benchmark information: “Like, give me the benchmark for that” (G1).

4.4.4 Participatory Design Study Findings

Such cases underscore the enduring challenge of embedding fairness into selection processes. While the question of whether to factor in markers of disadvantage or to concentrate solely on ‘task-relevant’ factors is a well-established debate in both selection committees and fairness literature [46], the primary difficulty often resides not in the question itself, but in the cautious and consistent application of any chosen approach. This is especially true when attempting to translate nuanced human considerations of fairness into algorithmic systems. Consequently, if algorithms do not adequately reflect these preferences for handling disadvantage, decision-makers will likely continue to observe discrepancies between algorithmically-driven selections and their own human-made judgments.

Exposure to SHAP explanations appears to have yielded useful insights when used as part of an ex-post decision-making process. In particular, participants appeared to find SHAP explanations useful for indicating when information may have been over- or under-used in selectors’ holistic review and in the algorithm itself. In this context, where decisions about individuals have already been made, and the organisation is looking into how to go about selecting its next cohort, the possibility of unwarranted trust in a particular output is not a concern. Rather, these explanations can be regarded as probes or provocations, helping decision-makers hone in on particular cases that highlight potential areas for improvement in decision-making, whether through changes in the model itself, or changes in human evaluation processes (e.g., placing greater or lesser weight on certain features). Thus, we can conclude that SHAP explanations may be useful ex post when trust in the primary output is not at issue. Interestingly, the wolf-husky example from Ribeiro et al. [150]’s original LIME proposal could be interpreted as being used similarly; using an explanation of an existing classification to guide changes in the model in future (e.g. by adding an edge detection step before classification, to ignore snow in

the background). In both cases, the explanation may reveal unwarranted reliance on (or lack of reliance on) a particular part of the feature space.

However, we also find that the SHAP-based waterfall explanations we provide, alone, lack the detail and presentation required. Practitioners desired additional context in the form of benchmarks, demographics, and selector-written comments; additional modes of interaction (e.g., the ability to change the importance of interview scores in the final algorithm score) with our explanatory materials; and points of comparison to clarify their investigation. This allays Miller [111]'s concern that post-hoc XAI methods might discourage explainees from engaging deeply with the facts of the task. Rather, in this case, the explanations served as a platform for explainees to seek additional information to inform their decisions.

4.5 Discussion

4.5.1 Implications

By delineating DSTs by the stage of the decision they inform, we can answer the question: Should we use post-hoc XAI methods?" separately for in-process and ex-post decisions. While we find them misleading, and thus dangerous, for in-process decisions, Section 4.3.4 indicates that these misleading tendencies are not limited to post-hoc XAI, and are rather a symptom of the practice of post-hoc justification more broadly. Furthermore, Section 4.4.4 indicates that certain ex-post use cases do not necessarily require that explanations appropriately modulate trust. Thus, post-hoc XAI methods might still inform ex-post decision-making (e.g., selection process refinement, evaluations of selector bias). This yields two direct implications for the XAI field:

1. While we reiterate caution around post-hoc justification of model outputs [9, 51, 73, 100, 111], we extend this caution from XAI methods to any form of post-hoc justification.

2. We qualify this caution in its application to ex-post decision-making. We encourage a field that has, in large part, moved on from post-hoc notions

of interpretability [13, 78, 88, 100] to engage with and identify ex-post applications for these tools.

4.5.2 The Anchor Problem

Suppose a farmer sees what they believe to be a sheep on a hill, and states "there is a sheep on that hill". Now, suppose this farmer sees a cleverly disguised goat, but that there is also a sheep on the hill, only invisible to the farmer. In this case, the farmer has a true belief ("there is a sheep on that hill") and has justification for it (the goat), but the justification is unrelated to the truth of the belief. In a seminal paper on Epistemology, Gettier [57] discusses this class of problem (now called 'Gettier Problems') and maintains that, despite the truth of the farmer's belief, that farmer does not know. In keeping with this tradition, Cabitza et al. [29] argue that, if an explainee is presented with a trust-inducing misleading explanation, even if that explanation induces trust in correct output, then the induced trust is misplaced.

In Section 4.3.4, we present what we believe is the most likely explanation for why Anchor explanations do not induce unwarranted trust: unlike SHAP and Confidence, these explanations might reveal concerns in the underlying model's local behaviour. However, other explanations exist. For instance, Miller [110] describes desiderata that make explanations well-suited to most explainees: explanations should be contrastive, counterfactual, selective, and social. While Anchor explanations are not social, they are contrastive, counterfactual, and selective. It may be that these explanations' beneficial effects on trust stem not from an ability to reveal concerns in the underlying model, but rather from these subjective desiderata. Another possibility is that the rule-based nature of Anchor explanations, often presented with precision and coverage metrics, might be perceived by users as more complex, less intuitively appealing, or inherently less certain than the feature attributions of SHAP or a simple confidence score. This perceived complexity or brittleness could lead to increased skepticism and a reluctance to fully trust the AI's output, especially when incorrect, thereby reducing unwarranted trust due to the explanation's format rather than its insight into model flaws. In either of

these cases, where the reduction in unwarranted trust is not due to the explanation faithfully revealing model issues, Anchor might still mislead [100].

4.5.3 Limitations and Future Work

One core limitation of our work relates to the choice of tasks. While *Salary*, *Credit*, and *Refinement* are closely related tasks, the distinction between in-process and ex-post decisions may be complicated by other distinctions between the three tasks. Future work should investigate this distinction in other contexts.

Another major limitation of our work stems from Section 4.5.2's Anchor problem. We recognise here that, though Chapter 2.3.3 and Section 4.3.4 give a compelling theory for the surprising results surrounding Anchor explanations, alternative explanations (such as the possibility that Anchors are not meaningful enough) exist, and we are unable to rule these out in our work. This calls for additional research or verification to provide better support for our novel findings with regard to Anchor explanations.

Finally, differences between the personalities of our online study and participatory design participants may limit the external validity of our results. Similarly, though we choose popular post-hoc XAI methods [13, 88, 151, 180], our choice of SHAP and Anchor limits applicability to other methods.

4.5.4 Conclusion

Miller [111] likens explanations provided by SHAP and related explanation systems to "Bluster", a hypothetical person that always gives a recommendation, even when unsure, and does their best to justify this. They note that, in the context of decision support, such a person is less valuable than "Prudence", who asks the decision-maker's opinion first and then provides feedback, as Bluster risks discouraging explainee engagement with the decision. Here, we distinguish between 'in process' and 'ex post' *decision stages*. We conclude that, while Miller [111]'s conclusion applies straightforwardly to the in-process stage, post-hoc XAI might still drive engagement and inform ex-post decisions, and urge that more be done to

identify and apply post-hoc XAI where it is useful. This chapter both hints at the significance of the distinction between in-process and ex-post decisions and hints towards the question: might other DSTs be more useful in process?

5

What Are Generative AI Detectors Good For? Evaluating and Implementing with the Decision Matrix[16]

Contents

[16]This chapter is based on a paper written in concert with Elías Hanno, Logan Gittelson, Reuben Binns, and Nigel Shadbolt. The paper is currently under review as: Neil Natarajan, Elías Hanno, Logan Gittelson, Reuben Binns, and Nigel Shadbolt. 2024. "What Are Generative AI Detectors Good For? Evaluating and Implementing with the Decision Matrix". Under review at CHI 2025.

5.1 Motivation

Chapter 4 elucidates an important distinction between in-process and ex-post decisions and finds post-hoc notions of interpretability suitable for supporting ex-post decisions, but not in-process ones. This chapter extends this work by conducting Action Research (AR) with the Programme A programme to codify this stage distinction alongside another axis of stakes. We apply this distinction to the problem of generative AI (GenAI) detection and again evaluate existing technology as a DST in the scholarship selection context.

5.2 Introduction

For comprehensive background on generative AI and academic integrity policies, see Chapter 2.3.4. Here we focus on the specific research gap this chapter addresses.

Since the rise of powerful generative AI models, students are increasingly using them to write essays, raising concerns about plagiarism [41]. While many GenAI users are caught by detectors, these detectors often fall short of their goal [41, 77, 99, 115, 168]. Essay-writers' ability to use GenAI has changed the role of essay-based teaching and evaluation, especially in competitive contexts like scholarship selection.

Many bodies of research seek to understand this new role. However, both theoretical and practical research has focused primarily on plagiarism detection and enforcement, inadvertently viewing the problem through a plagiarism lens [99, 115]. This focus on plagiarism neglects other problems GenAI has created for selectors.

We address this here with AR. We partner and "research with" [22] Programme A to understand the needs of their team of scholarship selectors (N=8; excludes authors) with regards to categorising and interpreting application essays and undertake this

research while supporting their Cycles X and Y. We identify "decision points" when the programme requires or desires to make a decision we might support with contextual information about GenAI usage (see Table 5.1). Through conversations with internal stakeholders [65], we identify two axes: stage and stakes, on which these decisions exist [23]; we use these axes to construct the Decision Matrix in Figure 5.2. We then evaluate three GenAI detectors on these decisions.

AR revealed several decisions that selectors desire to make surrounding GenAI. Though the literature focuses on the decision of whether or not to disqualify applicants using GenAI, Programme A dismissed this decision as irrelevant, as their application guidelines do not forbid the usage of GenAI. Instead, the programme expressed interest in decisions such as diligence, where the programme provides information about GenAI usage alongside other facts about the applicant to make a holistic decision about when and how to consider an applicant. We conceptualise these decisions in a Decision Matrix. The Decision Matrix framework identifies challenges that selectors and selection teams are faced with, reframing them in terms of "decision points". We then categorise them on two axes: stage and stakes. The stage axis captures the important distinctions between decisions made in the process of selection (in process) and decisions made after the primary *Selection* decision of "What cohort of people do we select?" has been made (ex post). The stakes axis, in contrast, captures the sensitivity (or severity) of the decisions. E.g., choosing to disqualify or select an applicant is a high-stakes decision, while choosing to assign extra staff to evaluate the truthfulness of an applicant's claims is a comparatively low-stakes decision. We then identify the properties desired from detectors so that they might support different decisions on different parts of our Matrix. In our evaluation of detectors for these decisions, we find that organisational needs are not met by current detectors, particularly concerning in-process decision-making. Our results suggest that detectors can be useful in aggregate analyses to support ex-post decisions. As a case study, we use one of our detectors, GPTZero, to support two decisions: *Partners* and *Pipeline* (described

in Table 5.1). We demonstrate two applications of detectors to support ex-post decisions. The flow of this study can be seen in Figure 5.1.

At a high level, this chapter has four major contributions:

1. We identify decision points facing scholarship and talent investment programmes regarding GenAI through AR.

2. We introduce the Decision Matrix framework for categorising decisions and evaluating the decision support capabilities of GenAI detectors.

3. We apply this framework to two detectors, GPTZero and Originality.ai, on Programme A's Cycles X and Y application data and determine their suitability for supporting specific decisions.

4. As a case study, We use GPTZero to support two decision points, *Partners* and *Pipeline*, in Programme A's context.

5.3 Methodology and Action Research

5.3.1 What is Action Research?

Action Research (AR) is a research philosophy that emphasises "research with, rather than on, people" [22]. Rather than one specific method, AR is best seen as a collection of related methods all embodying this ethos, usually to produce research contributions useful to the target group of people [102]. Among these are semiotic inspection [4, 40] and participatory design (PD) [20, 23, 61, 83]. AR is most often used in the context of social work, but can be applied across a variety of fields [43, 102].

In education, AR is often used in a classroom setting [108]. Venn-Wycherley et al. [174] argue that it is crucial in this setting to perform AR on both educators (teachers) and educatees (students), as failing to do so is liable to yield contributions useful to one group but not the other. While this holds for classroom settings, engagement across the stakeholder map is less feasible or desirable in scholarship selection. Unlike teacher and student, who share the common goal that the student

learn, selector and applicant are at cross purposes: selectors seek to choose the 'best' cohort of applicants (although they often disagree on what constitutes 'best'), while applicants seek to be included in the chosen cohort [16]. Thus, when elucidating the interests and desires of one group, the other will merely act as noise. (E.g., applicants who use GenAI to assist in writing their application will, of course, oppose using systems that monitor GenAI usage to disqualify applicants.)

AR is comparatively new to HCI [65, 102], but its methods and philosophies closely mirror longstanding pillars of HCI [65]. Much like PD and other HCI methods, AR seeks to democratise the research and design processes; unlike PD, AR extends beyond building solutions democratically, and sees learning through action as the ultimate research contribution [65]. For example, AR sees all parties become: "Co-investigators of, co-participants in, and co-subjects of...the project" [65]. Thus, research questions are formulated by and with participants, actions and interventions are designed by and with participants, and results are found by and with participants [65].

5.3.2 Our Action Research

In many scholarship selection contexts, organisations typically select groups of talented young people based on applications consisting variously of essays, videos, test scores, interviews, project results, group or individual activities, etc.. The organisation will begin by narrowing down the pool of candidates, often based primarily on test results, project results, or other easy-to-gather information. The organisation likely then compiles information from the application into short-form summaries of each applicant, complete with internally generated scores, and often even a recommendation. This recommendation is then often reviewed by a selection committee, who craft a cohort from the recommended applicants.

We engage Programme A in AR to investigate this selection context. In engaging this organisation in AR, several of the authors also functioned in supporting roles for the organisation's Cycles X and Y selection cycles. In addition to working alongside the selectors themselves, the authors were, at the time, a part of Programme A.

As with all AR, we first worked with participants (N=8; excludes authors) to identify research questions. We did this via a mixture of synchronous and asynchronous interactions. After this, we evaluated two GenAI detectors, GPTZero and Originality.ai, according to our research questions. Finally, we implemented one GenAI detector as an intervention. The flow of this study can be seen in Figure 5.1.

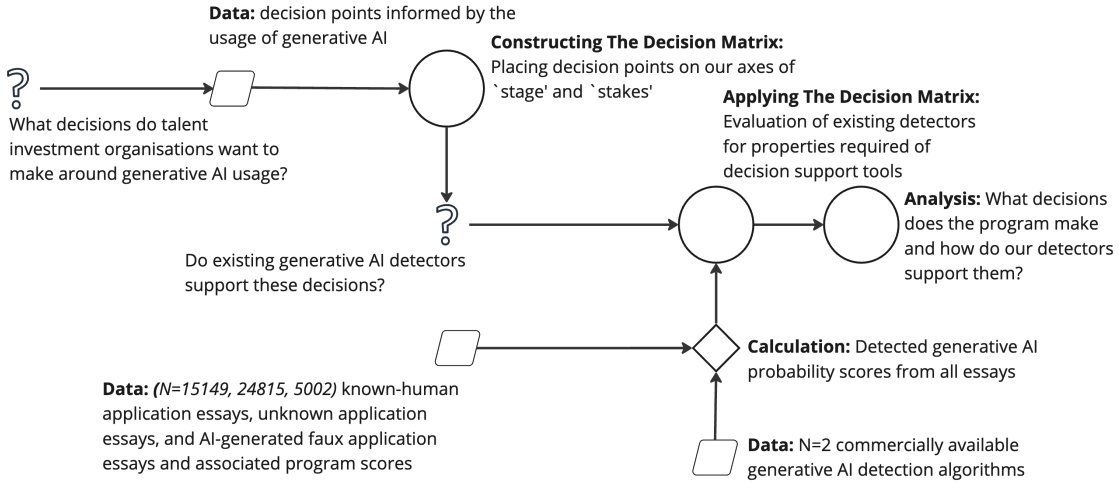


Figure 5.1: This figure describes the flow of our research.

Before our study, we obtained consent from all 8 participants to be included. Applicant essay data was collected by Programme A, who obtained consent to use these essays (anonymously) for research purposes. Participants also gave consent to be recorded, and to have these recordings stored on a secure server. All recording, transcribing, and data analysis was conducted on secure servers. Ethics review was performed by the University of Oxford's Central University Research Ethics Committee.

5.3.3 Positionality

Following Venn-Wycherley et al. [174], we state researcher positionality here. The research team is comprised of five researchers split between the United States and the United Kingdom; all researchers are men; four out of five researchers are ethnically White, while the fifth is South Asian; three researchers are affiliated with the University of Oxford and two are affiliated with an industry research foundation.

Table 5.1: Programme A discussed a desire to make several decisions concerning GenAI. These decisions and the information desired to support them are detailed in Chapter 2, but replicated here. Though the programme discussed disqualification, the programme has no application guidelines surrounding the use of GenAI and expressed no interest in disqualification. We include it here due to its ubiquity elsewhere in the literature [77, 99, 115, 168].

Decision Point	Decision Description	Supporting Information
Diligence	The programme makes holistic decisions about when and how to consider applicants.	Information about which essays (and which parts of essays) were written by GenAI; information about whether the GenAI-written passages are hallucinations.
Partners	The programme must determine whether to continue referral relationships, which encourage and support applicants.	Whether any referral organisations' affiliated applicants use GenAI disproportionately.
Pipeline	The programme decides whether to modify their application material or process.	Information about the usage of GenAI throughout the application pipeline.
Gameability	The programme decides how to modify their application material or process.	Information about how AI-generated essays are scored under the current application process.
Disqualification	A programme may decide to disqualify an applicant that violates their application guidelines.	Information about whether essays violate application guidelines around GenAI usage.

5.4 Constructing the Decision Matrix

We began from a starting point of "What do we do about GenAI?". Literature about GenAI elsewhere prompted early discussions to focus on: "Can we determine whether an applicant used GenAI to plagiarise?" but Programme A quickly discarded this; Programme A had no application guidelines forbidding GenAI usage, and hoped to accommodate "innovative uses of this powerful technology" in their selection process. From there, we quickly moved to "What decisions do talent investment organisations want to make around generative AI usage?". We then sought out

other decisions the programme wished to make in response to GenAI. Ultimately, we identified the decision points in Table 5.1; our final research question, then, is "Do existing generative AI detectors support these decisions?".

Notably, the decision to disqualify applicants for using GenAI (*Disqualification*) appears several times in the literature surrounding potential use cases for detectors [60, 77]. Thus, despite Programme A's dismissal, we list it alongside the different decisions discussed in Table 5.1.

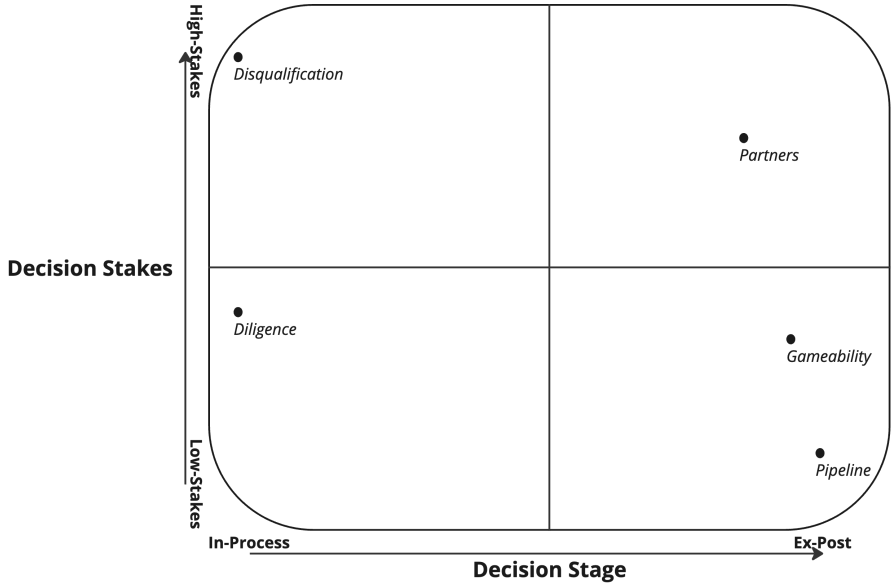


Figure 5.2: This figure places the decisions from Table 5.1 on the Decision Matrix, with axes of stage and stakes.

In conversation with Programme A selectors, we isolated two 'axes' on which the decision points exist: stage and stakes. Decision stages vary from entirely in process ('primary' decisions during the selection process) to entirely ex post ('secondary' decisions about future selection processes). Meanwhile, decision stakes vary from very low (e.g., a due-diligence decision to investigate further) to very high (e.g., a decision to disqualify an applicant). Figure 5.2 uses these axes to categorise the decisions in Table 5.1.

We also work with selectors to determine what type of properties a detection score should have to support each decision. For example, a disqualification decision may have stricter requirements than a diligence one. We lay out these desiderata on the same axes as Figure 5.2 in Figure 5.3.

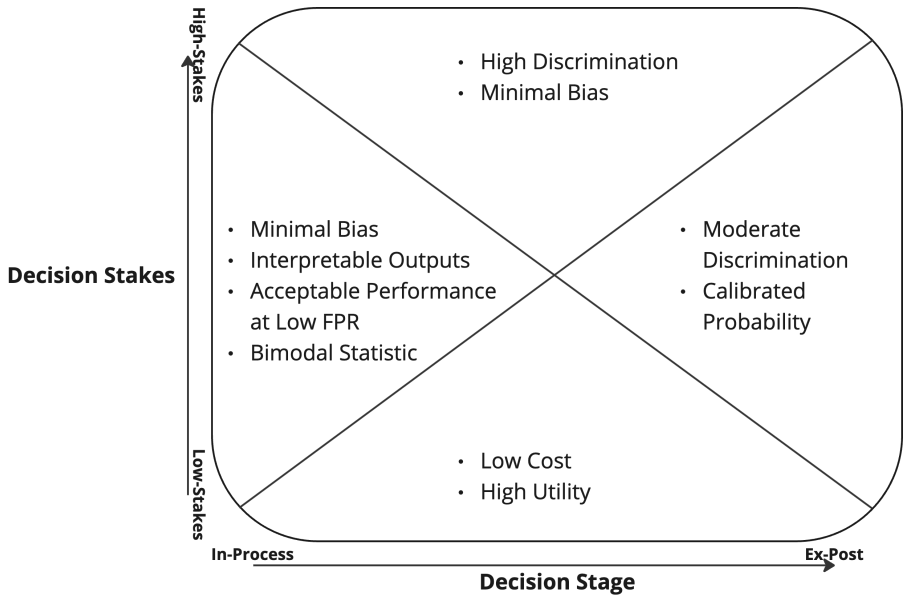


Figure 5.3: This figure places uses our Decision Matrix to understand the properties of GenAI detectors required (or, in the case of low-stakes decisions, desired) to support each decision.

5.5 Applying the Decision Matrix

5.5.1 Evaluating for Decisions

We seek to evaluate GenAI detection software for its suitability in supporting decisions across the Decision Matrix in Figure 5.2. To do this, we first evaluate detectors on the properties outlined in Figure 5.3; if a detector has the properties outlined in both low-stakes and ex-post decisions, it is likely to be useful in supporting decisions from this quadrant. However, although Programme A selectors indicated that these properties are necessary, we still seek to test their sufficiency. Thus, where the desiderata are satisfied, we apply the detectors to specific decisions from Figure 5.2. In particular, we select one decision from each quadrant (unless all detectors are deemed to lack the properties required for decision support in that quadrant). Finally, in case multiple detectors have sufficient properties, we compare detectors directly to make a recommendation about which method selectors should use for these use cases.

5.5.2 Data

5.5.2.1 Applicant Data

We use data from two of Programme A's application cycles, Cycle X and Cycle Y.

Applications for Cycle X were due before ChatGPT's public release, so we assume that these submissions were written without the use of GenAI [130]. Applications for Cycle Y were due after ChatGPT's public release, so GenAI tools were widely available and AI detection tools were already emerging [60, 81, 101]; we thus make no such assumption for these applications.[17]

Note: in this version of the thesis, specific cohort sizes, gender breakdowns, and regional grouping labels have been redacted to preserve programme anonymity. Approximate corpus totals are tens of thousands of essays per cycle. Applicants supplied gender identity and primary nationality, both grouped into programme-defined categories for analysis.

5.5.2.2 Synthetic Data

To obtain a set of known AI-generated essays, we generate a synthetic corpus using OpenAI's ChatGPT API responding to Programme A's Cycle X prompts. These sit alongside the human-written, applicant-submitted Cycle X essays to form a labelled corpus of known provenance. The Cycle Y corpus consists entirely of applicant-submitted essays, with unknown AI-content provenance. Specific corpus sizes have been redacted to preserve programme anonymity.

All of our synthetic essays were generated via OpenAI's ChatGPT API using GPT-3.5 [26]. More details on our prompts can be found in Appendix C.1.

5.5.2.3 Detectors

Despite the myriad of available GenAI detection tools, standardised comparisons of detectors are few and far between, but the benchmarks that do exist list similar models as leaders in accuracy across standard FPR thresholds. Dugan et al. [45]

[17] Several avoidance detection strategies (e.g., paraphrasers) have been proven to severely hamper state-of-the-art detection [77]. However, at the time of the relevant submission deadline, the efficacy of these detection avoidance strategies was not well-known. Thus, we assume these strategies were not widely employed.

introduce RAID, a standardised GenAI detection benchmark, and apply it to twelve popular models. Their results are mixed, but demonstrate a clear advantage for three detectors: the open-source Binocular model, and the commercial GPTZero and Originality.ai models [45]. Verma et al. [175] compare GPTZero, DetectGPT, two baselines, and their own model, Ghostbuster. Their results are similarly mixed but, in a variety of scenarios, GPTZero or Ghostbuster variously perform best [175]. Programme A reached out to both GPTZero and Originlity.ai, and both offered access to their models for research purposes.

To calculate scores, we use the API under default settings for both GPTZero and Originality.ai. Note that we calculated our GPTZero-based detection scores at the time of the corresponding application cycle, while the Originality.ai scores were calculated more recently for this thesis. Thus, a more recent version of GPTZero's model may have since been made available; our results apply only to the version of GPTZero we used at the time. This yielded various statistics from each detector for each essay, but we are primarily interested in the overall likelihood statistic from each.

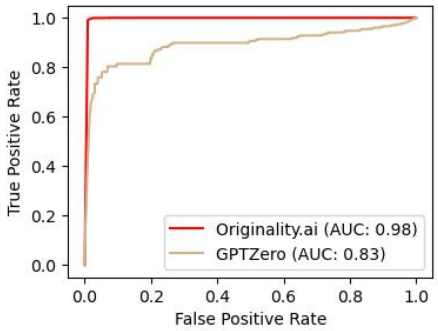


Figure 5.4: This Receiver Operating Characteristic (ROC) curve shows the performance of each detector on our data of known providence (Cycle X submissions and ChatGPT responses to Cycle X prompts). While GPTZero's ROC curve has a moderate area under the curve (AUC), Originality.ai's ROC curve has a very high AUC.

5.5.3 Results

5.5.3.1 Both Detectors Possess the Properties Desired for Low-Stakes Decision Support

We identified no properties required to support low-stakes decisions, but list several desiderata that we consider more a matter of utility than of necessity. In particular,

detectors should have:

1. Low cost

2. High utility

We measure cost as the price per essay evaluated at the highest tier of subscription service available from each detector in Table 5.2. We measure utility according to ROC AUC.

Table 5.2 shows that both detectors have a low cost, with GPTZero having a slightly lower cost.[18]

Figure 5.4 demonstrates high utility from both detectors, though Originality.ai has a higher ROC AUC than GPTZero. Overall, both detectors satisfy the desiderata specific to low-stakes decisions.

Table 5.2: This table displays the estimated per-essay costs of each detector; both detectors are deemed sufficiently cost-effective.

Detector	Cost per Essay
GPTZero	$0.023
Originality.ai	$0.06

5.5.3.2 GPTZero Possesses the Properties Required for High-Stakes Decision Support

We identified two properties required to support high-stakes decisions. Detectors should have:

1. High (predictive) discrimination

2. Low bias

We measure discrimination according to whether the detector discriminates positive and negative cases (see Table 5.2). We measure bias as the difference in FPR between demographic groups.

[18]As we used both detectors for research purposes, we reached out to the companies involved and were given research access. These cost estimates are based not on our access, but on public information assuming the highest tier of subscription service available from each detector.

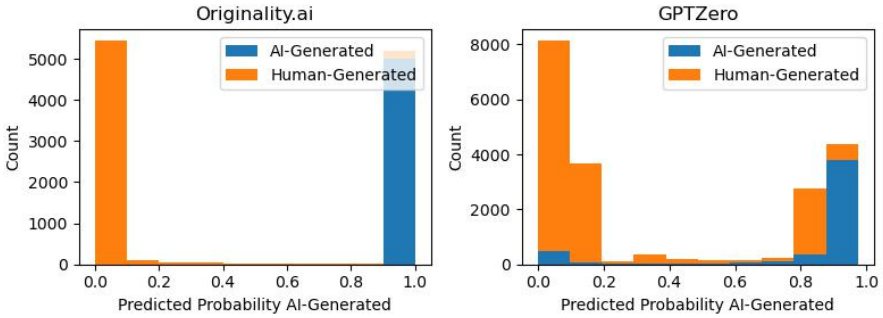


Figure 5.5: These histograms demonstrate the high bimodality and predictive discrimination of both detectors' outputs.

As can be seen in Figure 5.5, both detectors' outputs are close to either 0 or 1, though Originality.ai has far higher discrimination than GPTZero. This is reflected in the confusion matrices in Figure 5.6. As can be seen, both output statistics discriminate well between positive and negative cases. However, while GPTZero discriminates between positive and negatives, Figure 5.6 shows that Originality.ai has less error of both types.

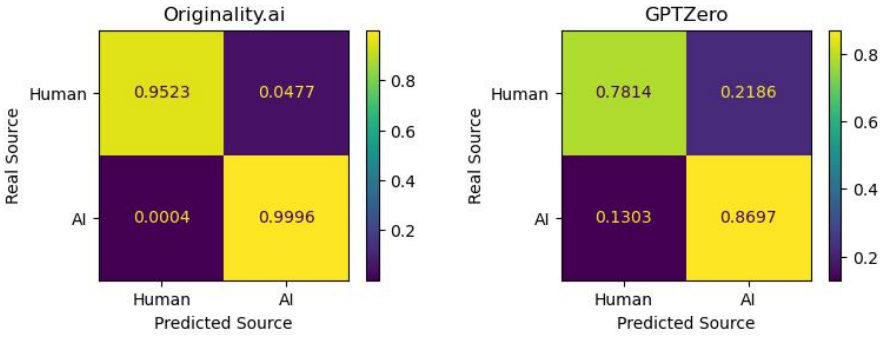


Figure 5.6: These confusion matrices again demonstrate the high predictive discriminative capabilities of both detectors' outputs.

We measured per-demographic FPRs for both detectors in the human-written Cycle X submissions. Specific per-region FPR values and ANOVA statistics have been redacted to preserve programme anonymity. The high-level finding: GPTZero exhibited higher overall false-positive rates but more uniform performance across demographic groups, while Originality.ai showed lower overall FPRs but substantial regional variance. We considered GPTZero suitable for ex-post decisions across our applicant pool; Originality.ai's regional bias rendered it unsuitable for that use.

5.5.3.3 Both Detectors Possess the Properties Required for Ex-Post Decision Support

We identified two properties required to support ex-post decisions. Detectors should have:

1. Moderate Discrimination
2. Calibrated Probability[19]

We measure discrimination as in Section 5.5.3.2. We measure statistic calibration using calibration curves in Figure 5.7.

We have already determined in Figure 5.4 that Originality.ai has particularly high predictive discrimination, while GPTZero has moderate predictive discrimination suitable for ex-post analyses.

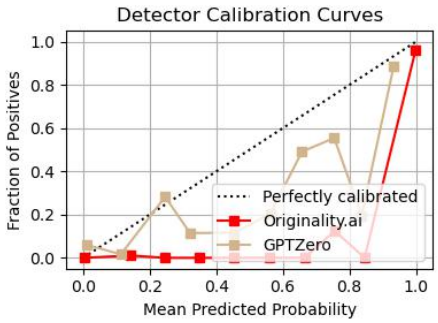


Figure 5.7: These calibration curves demonstrate the bimodality of both detectors' outputs. As can be seen, both output statistics fall far below the calibration curve; these output statistics are not calibrated on our data.

In aggregate analyses, we desire to reason about the probability that a particular essay was generated by AI ($P(AI)$), or even the expected number of AI-generated essays in a given group ($E(P(AI))$). This yields convenient general properties, e.g., $E(P(AI))$ is just the mean $P(AI)$ within a given group. Figure 5.7 provides a calibration curve for both detectors and demonstrates a comparative lack of calibration in both cases. In both cases, output statistics fall far below the calibration curve, indicating a bias towards extrema. In effect, this entails that we should

[19]Calibrated statistics are ones whose distributions are probability-like in expectation.

not treat these output statistics like probabilities. However, so long as we have a large provenance of labelled data (which we do in the form of tens of thousands of applicant-submitted and several thousand ChatGPT-generated Cycle X essays), we can calibrate an uncalibrated statistic to achieve a more probability-like output. In our case, we calibrate our statistics on our data by applying a monotonic transformation to ensure that, within any subset of our body of essays, the mean predicted probability aligns with the fraction of positive cases.

Thus, we conclude that both calibrated statistics possess the properties of probabilities that we would require for use in ex-post analyses. We caution other organisations against using these detectors for these purposes without first calibrating their output statistics on data of known provenance.

5.5.3.4 Neither Detector Possesses the Properties Required for In-Process Decision Support

We identified four properties required to support in-process decisions. Detectors should have:

1. Minimal Bias
2. Interpretable Outputs
3. Acceptable Performance at Low FPR
4. Bimodal Statistic

We measure bias as in Section 5.5.3.2. We reason about model interpretability. We measure acceptable performance at low FPR based on TPR rates at an FPR of %1. We measure bimodality using the histogram in Figure 5.5 and the calibration curve in Figure 5.7.

Recall that, we observe bias in both outputs and conclude that, while GPTZero is sufficiently balanced across demographic groups, Originality.ai's outputs display too much regional bias for our purposes.

We also note that GPTZero provides interpretability in the form of local-level scores that might direct a human overseer's attention to particularly problematic phrases or sentences. Originality.ai, in contrast, only provides a single overall statistic.

Figure 5.5 demonstrates the bi-modality of both detectors. Though it is clear that Originality.ai's output is more bimodal, we consider both detectors sufficiently bimodal for our purposes.

Table 5.3: This table displays TPRs at %1 FPR.

Detector	TPR
GPTZero	%36.0
Originality.ai	%98.9

Finally, to analyse performance at low FPR, our organisation set the acceptable FPR at %1. We see a TPR of %36.0 for GPTZero and %98.9 for Originality.ai in Table 5.3. While GPTZero's output statistics are interpretable and display limited bias, it fails at low FPR rates.

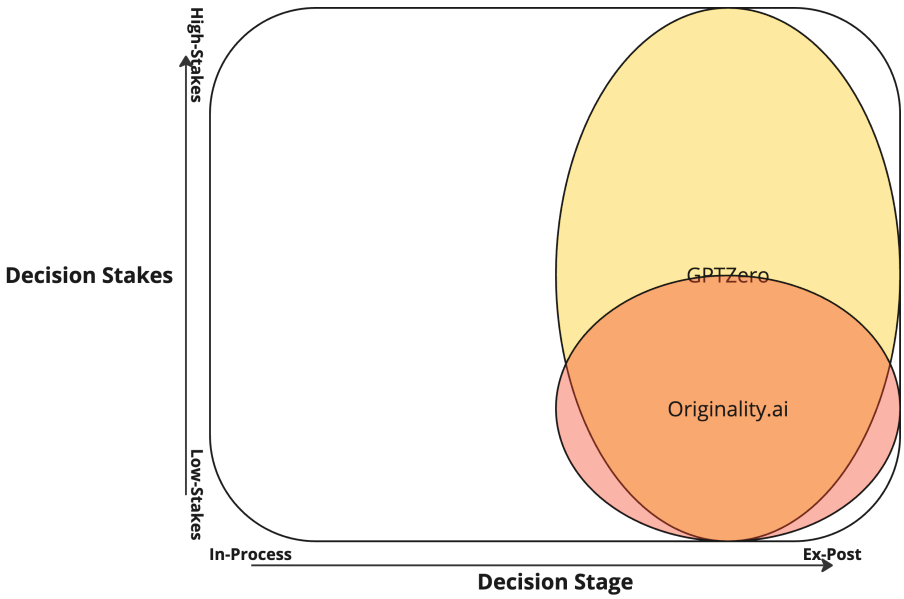


Figure 5.8: This figure demonstrates the results of our application of the Decision Matrix, marking the use cases we consider suited for each detector.

5.6 Analysis of *Pipeline* and *Partners*

5.6.1 Low-Stakes Ex-Post: Analysing Overall GenAI Usage in a Recent Application Cycle (*Pipeline*)

We focus our subsequent analysis primarily on the potential use of GenAI by applicants in Cycle Y. Seeking to avoid the disproportionate effects induced by GPTZero's heterogeneous biases, we focus primarily on within-group changes. We note here that the mean probability of an essay being AI-generated within a corpus is exactly the expected proportion of AI-generated content within that corpus. Thus, we test for changes in mean $P(AI)$. As we have previously confirmed that GPTZero is suitable for these analyses, we use our calibrated GPTZero score going forward.

We find a small overall increase in mean $P(AI)$ between cycles, but the magnitude is limited and within-region changes do not align with a simple increased-GenAI-use interpretation: some regions show statistically significant decreases that pre-date GenAI's wide availability, suggesting cohort-level variance or applicants' use of AI-detection tools to evade flagging. Specific per-region findings have been redacted to preserve programme anonymity. In any case, this analysis surfaces interesting discrepancies demanding further interrogation in future cycles but does not demand the programme alter its application material or process.

5.6.2 High-Stakes Ex-Post: Analysing GenAI Usage Across the Programme's Referral Organisations (*Partners*)

As a final analysis, we evaluate whether GPTZero detects any suspicious patterns in the essays associated with the programme's referral organisations, who refer and support applicants to the programme. The organisation partners with 'referral organisation' organisations that encourage applications and these are attributed to referral organisations based on the custom links that applicants use to reach the programme's website, as well as questions in the application about how applicants learned of the programme. Evidence that any of these referral organisations used GenAI to create large volumes of applications would warrant further investigation before continuing affected partnerships.

In analyzing GenAI usage associated with referral organisations, we consider the proportion of essays classified as 'AI-Generated' rather than simply averaging the raw detector scores across all submissions for a partner.[20] For this analysis, we use a cutoff of 0.5 on our calibrated predictor, meaning that essays flagged as 'AI-Generated' are more likely than not to be so. This threshold has a TPR of 76% and an FPR of 4%. We limited the analysis to the partners with the most essays and have hidden individual partners' identities.

We found that only one of the referral organisations studied was associated with essays identified as AI-generated at a rate meaningfully different from the overall pool, and that partner was associated with *fewer* flagged essays than average. We attribute this to multiple-hypothesis-testing chance and to underlying demographic correlations identified in earlier-cycle analyses. Specific per-partner statistics, regional attributions, and partner identifiers have been redacted to preserve programme anonymity. We find no evidence of widespread GenAI use among the programme's referral organisations, supporting the organisation's referral approach and choice of partners.

5.7 Discussion

5.7.1 Implications for the Programme

The programme we work with has several use cases for GenAI detection, and we have found that GPTZero is suitable for the ex-post use cases, both high- and low-stakes. However, the programme's use of GenAI detection for in-process decisions is not yet feasible, as, although Originality.ai has very high accuracy, it suffers from a lack of interpretability and its FPRs exhibit large heterogeneous biases across demographics, especially regions. We recommend that the programme continue to use GPTZero for ex-post decisions, but that it seek out a new detector or design its own if it wishes to make data-driven, in-process decisions surrounding GenAI usage. A summary of our recommendations can be found in Figure 5.8.

[20]Averaging can obscure the impact of a few essays with very high AI-likelihood scores if the majority score very low, making it harder to detect potential issues. By using a classification threshold, we can more directly identify the incidence of essays deemed likely to be AI-generated.

5.7.2 Implications for Other Programmes

By identifying decision points, placing them on the Decision Matrix, aligning these desired and required properties, and then testing GenAI detection tools for the relevant properties, programmes can determine the suitability of detectors in supporting and informing decision points. If a programme's decision points and desired properties align closely with Programme A's, they may find that they consider the same use cases for GPTZero and Originality.ai, and have no need for our framework. But if the GenAI detection landscape changes in response to new developments [72, 77, 99, 101, 115, 130, 173], or if programmes' priorities do not align with those of Programme A, programmes should use the Decision Matrix framework to replicate the analysis done here.

5.7.3 Implications for the Field

Our results suggest that, while GenAI detection tools address some issues faced by scholarship selectors, deeper involvement with these institutions is key in designing technology to meet the specific needs of these selectors. Currently, a narrow focus on academic integrity, plagiarism, and decisions to censure essay writers limits broader discussions about how GenAI should be integrated into academic workflows. While selectors shift towards embracing GenAI as a tool rather than a threat, the field of HCI is lagging behind.

Our AR process raises the question: should it matter whether GenAI was used at all? Rather than aiming to detect AI-generated plagiarism, selectors are perennially concerned with determining whether an applicant's submission indicates their aptitude for the programme; the need for detector-based decision support, in all cases, is to support decisions impacting that ultimate determination. GenAI tools pose problems to selectors' abilities to make that determination with contemporary essay-based assessment, but outright bans and disqualification [170] unenforceable with current technology offer no solutions. A shift in assessment methodology, then, is a practical and desirable alternative.

HCI and GenAI detection can enable this shift through respectful design [172] by working with selectors to understand their needs and support them (e.g., by improving assessment design or by building digital feedback mechanisms). Institutions, meanwhile, may wish to design essay prompts that encourage or even require applicants to use GenAI in a meaningful way. In such cases, the role of detectors would shift from merely identifying AI-generated content to evaluating how well applicants have leveraged these technologies.

5.7.4 Ethical Implications

The Decision Matrix framework developed throughout this chapter reveals a variety of desiderata for the usage of GenAI detectors in application essay settings. As can be seen in Figure 5.3, the decision to disqualify a candidate demands much of detectors. Indeed, even when detectors possess all of the desired properties, taking automated adverse action against applicants is ethically fraught [93]. However, when opting to make no decisions about applicants using this technology, these demands fall away.

In practice, our research sits between these extremes. The process of holistic review sees organisations incorporate a mass of disparate information into an opaque decision-making process. Hirschman et al. [67] note this lack of transparency as a benefit of holistic review insofar as it shields institutions from regulators. But while this may offer some legal insulation to the holistic review process, no such moral insulation exists. In-process algorithmic decision support, even for low-stakes decisions such as *Diligence*, still bears a high moral burden [93].

5.8 Limitations and Future Work

One participant highlighted: "There are two problems here: Did this applicant use [generative AI]? And if so, is this essay based in fact?". This points to a key limitation of our evaluation of detectors: while we could determine whether text was AI-generated, we had no basis for evaluating the truthfulness of AI-generated text. As GenAI is known for its hallucination [3], frequently yielding convincing falsehoods, this represents a significant limitation. While this distinction is ultimately irrelevant

to those who consider a GenAI usage plagiarism [170], decisions such as *Diligence* would be well-informed by an understanding of how GenAI was used. Future work should investigate detecting the nature of GenAI usage (writing, editing, etc.) and then determining the truthfulness of GenAI-written text.

As we select only two detectors, our results as applied to these detectors may not apply to others. Furthermore, as the field of GenAI detection moves so rapidly, our results from Cycles X and Y may not even apply to future cycles. For this reason, we develop the Decision Matrix framework to support ongoing evaluations of detectors in response to new developments such as paraphrasing tools and hybrid human-GenAI writing processes [77]. Future work should use this framework in re-evaluating detectors in response to these new developments.

We deliberately do not engage applicants in our process. Venn-Wycherley et al. [174] argue that, when conducting human-centred research in a classroom setting, it is important to gather perspectives of both educators and students. Here, we conduct AR centred on scholarship selectors, but we omit the perspectives of their young decision subjects. Unlike in the classroom context, the evaluation context sees an adversarial relationship between the educational institution and its target population – while the scholarship seeks to select only the most well-fit candidates, each candidate seeks to be selected, and therefore to make themselves seem most fit. Thus, in making the selector "Co-investigators of, co-participants in, and co-subjects of" our research [65], we necessarily exclude the perspectives of their decision subjects. Future work should seek to engage these young decision subjects in this context and may explore concepts like the essayist's sense of authorship, the line between writing and editing, essayist thoughts on plagiarism, and applicant perceptions of programme decisions driven by AI.

5.9 Conclusion

In summary, this research examines the various decisions that may arise when scholarship selection organisations consider the problems posed by GenAI in practice, emphasising the need for tools designed to support decisions besides

simply disqualifying applicants. Our findings reveal that, although state-of-the-art detectors may be unsuitable as automated disqualification tools, they can be used as-is to support "integrating technology, education, policy reform, and assessment restructuring" [140] and support ex-post decisions organisations may wish to make. By engaging in action research, we catalogued real decisions scholarship selection organisations seek to make in response to the problems posed by GenAI usage. We then worked with them to develop the Decision Matrix, which serves as a tool for selectors to evaluate GenAI detectors on their data. As we move forward, we call for a broader view of the purpose of GenAI detection, and for a restructuring of what it is to learn and assess in an era so heavily influenced by easy access to GenAI tools. We also call for more research into the Decision Matrix, specifically examining the support of the ever-elusive in-process decision points; Chapter 6 seeks to support one such in-process decision point.

6

"Diversity is Having the Diversity": Unpacking and Designing for Diversity in Applicant Selection[21]

Contents

[21]This chapter is based on a paper written in concert with Sruthi Viswanathan, Reuben Binns, and Nigel Shadbolt. The paper is currently under review as: Neil Natarajan, Sruthi Viswanathan, Reuben Binns, and Nigel Shadbolt. 2024. "'Diversity is Having the Diversity': Unpacking and Designing for Diversity in Applicant Selection." Under review at CHI 2025.

6.1 Motivation

Chapters 4 and 5 both reveal flaws of existing algorithmic support tools in their applications to decision points facing scholarship selection organisations: these tools lack the properties required to support in-process decision-making. In this chapter, we select a family of in-process decision points of particular interest to the literature: ensuring diversity in selection. This chapter, thus, seeks to understand what diversity is and how to support its consideration.

6.2 Introduction

For comprehensive background on diversity as a societal value, historical injustices in technological systems, and approaches to measuring diversity, see Chapters 2.2 and 2.3.1. Here we focus on the specific research gap Chapters 6 and 7 address.

Processes for selecting people for jobs, universities, prizes, and other opportunities have often failed to reflect the diversity of their actual and potential candidate pool. Recognising this, various sectors have in recent years shifted towards recognising and promoting diversity through the establishment of a variety of related norms: DEI (Diversity, Equity, and Inclusion), EDI (Equality, Diversity and Inclusion), JEDI (Justice, Equality, Diversity, and Inclusion), DEIB (Diversity, Equity, Inclusion, and Belonging), etc. [70, 113, 143]. These norms are frequently operationalised through changes to application, evaluation, and decision-making procedures designed to result in greater representation of different demographic groups in final selection decisions [143]. Such efforts are in part a response to widespread societal concerns about racial, gender, and other injustices, but have

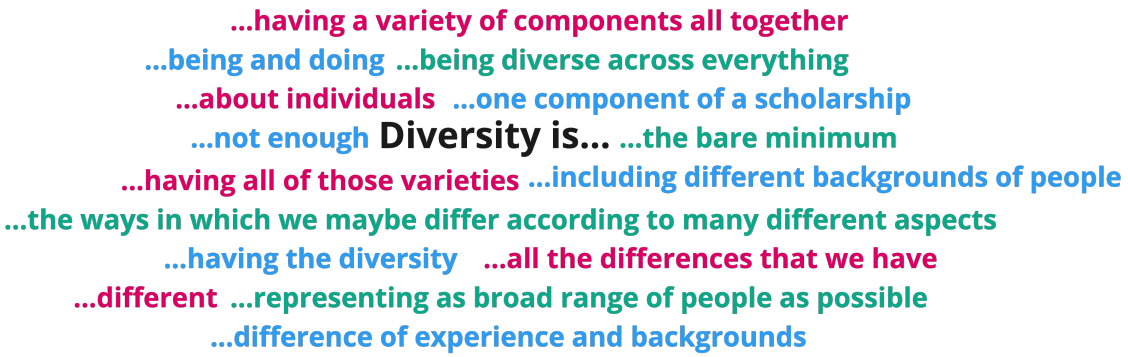


Figure 6.1: This figure shows participant codes defining what "diversity is". In this chapter, we seek to answer: what do they mean and how do we design for that?

also frequently been justified in economic terms by evidence suggesting that diverse teams perform better than homogeneous ones on a variety of tasks [42, 128, 134]. Meanwhile, the concept of diversity itself has become swept up in 'culture war' discourse, criticised by right-wing commentators as part of a sinister 'woke' agenda, and by progressives as mere window dressing that fails to meaningfully address deeper societal injustices.

For practitioners on the ground, however, the question of how to meaningfully measure and promote diversity is a practical challenge. The proliferation of software for hiring and selection presents both complications and opportunities [93]. AI-driven tools may discriminate [31, 91, 98], but they could also help mitigate human biases and increase diversity [8, 166, 182, 184]. This context challenges Human-Computer Interaction (HCI) research to improve diversity in selection without amplifying existing inequities.

But to improve diversity, we must first understand it. We undertake two studies to do so. In Study 1, we conducted 15 one-to-one interviews with selection practitioners (selectors) from global scholarship programmes to understand how they define and operationalise diversity. Our inductive thematic analysis revealed the ad-hoc nature of current practices and surfaced three distinct definitions of diversity: as 'different perspectives' in the same room; as 'representativeness' of a target population; and as 'contextualising applications' with information about an applicant's relative privilege. We conclude that technological interventions should

first identify which definition(s) of diversity they aim to promote, and we construct the *Diversity Triangle* (Figure 6.5) as a conceptual guide.

In Study 2, we developed six design prototypes based on the Diversity Triangle and presented them to participants in participatory design workshops. The prototypes included tools for visualising cohort representativeness, measuring entropy (the average number of in-group differences), and assessing individual applicant (dis)advantage. They were presented to participants via participatory design workshops [190]. The workshops revealed that while quantitative tools are essential, selectors rely on their own idiosyncratic lenses and qualitative assessments to navigate diversity considerations.

This research contributes to understanding how data-driven tools can support diversity in selection. Our contributions are:

1. Three practitioner-led definitions of diversity uncovered through inductive thematic analysis.

2. The Diversity Triangle, a framework for categorising diversity-related themes.

3. Design recommendations grounded in participatory design for building diversity-supporting tools.

More broadly, this work demonstrates that by providing structured, data-supported approaches to diversity, organisations can better navigate the complexities of DEI (EDI, JEDI, DEIB, etc.). While differing in some respects, we believe these implications will generalise from scholarship selection to various other talent identification contexts, including recruitment for jobs and admission to universities. By helping these organisations achieve their desired outcomes in selection processes, we aim to ultimately contribute to and help build a more diverse society.

6.3 Experimental Design

6.3.1 Our Studies

Our research is broken into two components: 15 one-to-one interviews with scholarship and talent investment selectors and two participatory design workshops with a subset of the 15 selectors. The relationship between these components is shown in Figure 6.2 and details on Studies 1 and 2 can be found in Sections 6.4.1 and 6.5.2, respectively.

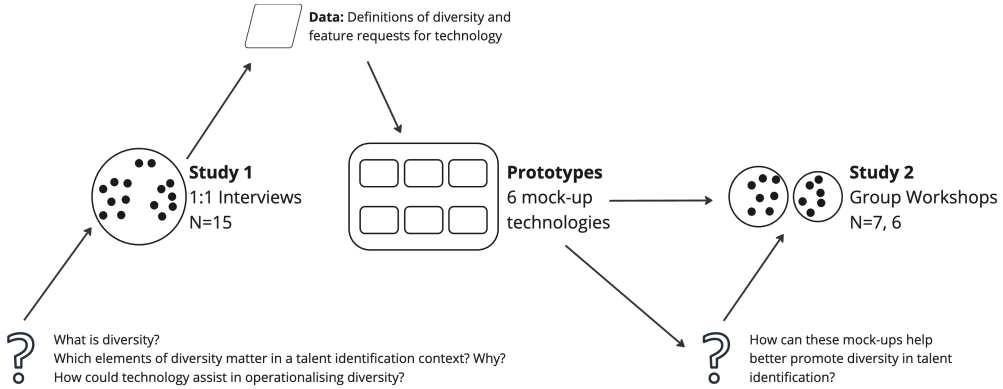


Figure 6.2: This research begins with 15 interviews seeking to understand what selectors mean when they talk about diversity and how to support that, followed by two scenario-speed-dating activities where these selectors test several prototypes built based on the interviews.

In the first session, we seek to ascertain how these selectors understand diversity, how they operationalise it in processes for selecting talented applicants, and how they envision using technology to assist them in that process. In the second session, we show these selectors six prototypes built in response to the interviews, then we aim to collaboratively design tools that can help them better consider diversity in selection.

6.3.2 Positionality

Following Venn-Wycherley et al. [174], we state researcher positionality here. All authors endorse diversity as a societal and organisational value (while sympathetic to the critiques of Ahmed [1] and Warikoo [179]); we thus contend that improving organisational capability to consider diversity is generally a positive development.

The research team is comprised of three men and one woman; ethnically, two researchers are South Asian, while two are White; researchers represent three different primary nationalities; all researchers are affiliated with the University of Oxford.

6.3.3 Participants

We engaged selectors (N=15) from two scholarship and talent investment programmes. These participants are numbered P1-P15.

Before either study, we obtained informed consent from all participants to be included in both studies. All participants were given the option to recuse themselves from Study 1 at any point until publication, but, as Study 2 is a group workshop (and as we do not do participant-level attribution), participants were asked to recuse themselves before this second study. Two participants recused themselves before Study 2 but gave leave to be included in Study 1. Participants also gave consent to be recorded, and to have these recordings stored on a secure server. All recording, transcribing, and data analysis was conducted on secure servers. Ethics review was performed by the University of Oxford's Central University Research Ethics Committee.

6.4 Study 1: Interviews and Thematic Analysis

6.4.1 Methodology

Our interviews aim to answer three questions in each organisational context:

1. What is diversity?
2. Which elements of diversity matter in a selection context? Why?
3. How could technology assist in operationalising diversity?

In answering our first set of research questions, we follow Braun and Clarke [23]'s methodology for reflexive thematic analysis. We first conduct 45-minute semi-structured interviews with 15 selectors. In each interview, we first ask general questions about their selection methodology; we next ask specifically about diversity

and its role in selection; we move on to Knapp et al. [83]'s 'crazy 8s' exercise, where participants give eight feature requests in eight minutes; we conclude with a 'magic app' exercise inspired by Blythe [20]'s design fiction, where participants more thoroughly detail their ideal app. A question-by-question protocol for these interviews is supplied in Appendix B.2.

The lead author interviewed participants and then transcribed and anonymised the interviews. The lead author and another author (who wasn't present during the interviews) then independently 'open-coded' each anonymised transcript to mitigate bias, looking for anything relevant to our research questions. The researchers met six times to discuss their open codes, then shared these codes with the remainder of their research team across four meetings; the researchers grouped codes into 6 themes and 18 subthemes after consensus was reached. These themes are detailed in Table 6.1 and described in Section 6.4.2.

6.4.2 Themes

6.4.2.1 Why Diversity?

A central theme of our investigation revolved around the question of why diversity matters. To this end, we asked questions such as: "What is diversity?" and "Why does diversity matter?". Answers to: "What is diversity?" are visualised in Figure 6.1; as hypothesised, these answers are vague and uninformative. Interestingly, though, answers to: "Why does diversity matter?" informed more specific definitions. We were able to cluster these more specific definitions into three central subthemes: 'different perspectives', 'representativeness', and 'contextualising applications'. These are listed as subthemes of 'Why Diversity' in Table 6.1.

Different Perspectives 8 out of 15 participants mentioned that diversity was important because it brought different perspectives into the same room. This was seen as important for a few related reasons, e.g., the ability to see problems from different angles and the ability to make better decisions. Several participants referred to the: "Benefits of diverse perspectives" (P1). One said, when discussing

Table 6.1: Our three central subthemes all speak to the question "Why diversity?". Other themes and subthemes reflect types of diversity, concepts intertwined with diversity, and other considerations scholarship programmes must weigh against diversity desires.

Themes and Subthemes	
Why Diversity?	**Types of Diversity**
Different perspectives	Socioeconomic
...in the same room	*parental income*
Representativeness	*parental education*
...of a general population	*generational wealth*
...of the eligible population	Sex, gender, and sexuality
...of the applicant population	*sex*
...of a target population	*gender identity*
Contextualising applications	*sexual orientation*
	Geography
	nationality
	the 'Global South'
	region
	Race
	international categorisations of race
	Types of thinking
	subject area interest
	personality type
	core beliefs
	problem solving approaches
	political views
Operational Risks and Considerations	**Fairness and Bias**
Outreach	Fairness
Support	*...to the applicants*
...during the application process	*...to the world*
...after selection	Bias
Selectors	*measurement bias*
Applicant fraud	*decision-makers' bias (prejudicial)*
	decision-makers' unique perspective (probative)
Scholarship Goals	**Merit**
Impact	Performance relative to disadvantage
...on all applicants	Measurement
...on the selected scholars' performance	Performance
...on the selected scholars' opportunities	
...by the selected scholars on the world	

their personal experience working with winners in a talent investment programme, that there is: "Magic happening with lots of...diverse perspectives in the room" (P14). People's experiences were particularly relevant here. As one participant writes: "You want to have diverse perspectives from people who look different with different experiences" (P2).

Representativeness 9 out of 15 participants spoke of 'representativeness'. This, we observed, was often spoken of in relationship to a larger population. Most frequently, participants spoke of the importance of having a cohort that was representative of the 'eligible population'. I.e., one participant said: "[You want] a community which is representative of where you are selecting young people from" (P6). Others spoke of this in broader, more general terms: "[You want] as broad a range of people as possible" (P15). Participants identified the importance of building a cohort that variously: "Reflects the population of the countries" (P15) and is "More representative of the national population than the STEM field already is" (P10). Others talk about the representation of a particular target population, i.e.: "Representation...because that gives you insight for the people that you're trying to serve....you have to be....reflective of your market" (P9). Finally, selectors discussed the importance of representing an applicant population: "[We want] a cohort that is representative of the pool" (P10).

Contextualising Applications 'Contextualising applications' was often spoken of by 6 out of 15 participants, most often in individual terms, speaking of identifying particular applicants in need of support, and then offering them a 'boost' in the form of said support. I.e.: "Identify those talents and specifically boost up people who are in need of support" (P5). One participant identified 'boosting' as a key metric for the programme: "We need to know that...we have some level of...impact here, and...if all we're doing is supporting someone who is already on an amazing trajectory and then maybe that means we're not altering their trajectory at all. That's a question of efficiency of our dollars" (P8).

In some cases, the need for support or boosting was identified with under-represented or disadvantaged demographic groups. One participant said: "The focus on gender has been to give the sex that has had the least opportunity the opportunity in this programme" (P9).

6.4.2.2 Types of Diversity

Another central focus of our investigation was on different types of diversity. We asked participants to break down their understanding of diversity into different elements and to discuss why these elements were important. We found that participants identified a wide range of different types of diversity, which we clustered into subthemes. These subthemes are listed as subthemes of 'Types of Diversity' in Table 6.1.

Notably, in addition to the standard demographic categories commonly considered 'demographic diversity', participants identified a wide range of other types of diversity commonly termed 'cognitive diversity' [133]. These included 'subject areas of interest', 'personality type', 'core beliefs', 'problem-solving approaches', and 'political views'.

Socioeconomic All 15 participants identified socioeconomic diversity as particularly important in the context of a talent investment programme. One participant said: "Socioeconomic [diversity] is probably the most important" (P1). Another said: "Socioeconomic background is number one from my perspective" (P5). This was identified as particularly important for several reasons. Participants stated: "Because right now the SAT, for example, is more highly correlated to socioeconomic status than it is to anything else" (P5), and: "It's a scholarship scheme, so I think it should be for kids who cannot afford normally the fees at the university" (P7).

Outside of the standard categorisations by income and wealth, participants also identified: "Familial education level" (P5) or "Socioeconomic backgrounds" (P10) as a particularly important metric for understanding socioeconomic diversity in a scholarship context. This suggests that historical socioeconomic status is considered alongside present socioeconomic status. One participant noted that

socioeconomic status varied in both meaning and measurement from country to country: "For example, in Columbia, there's a whole society to organise on a 1 to 7 scale for socioeconomic status" (P5).[22]

Sex, gender, and sexuality While all 15 participants noted some manner of sex, gender, and sexuality diversity as important, participants disagreed on the relative emphasis that should be placed on each. One participant noted: "[Sex] is important. I think it will get diluted if we focus on identity gender because....the purpose of diversity on the gender aspect was to make sure that [men and women] were getting equal opportunities" (P9). Another noted that, while sexual identification diversity was important in other contexts, they "Wouldn't select for that" (P1) in this context. Others listed 'sexual orientation' and 'gender' as important metrics for understanding diversity in a scholarship context. However, save for the participant who noted the distinction between sex and identity gender, participants expressed reluctance to discuss the relationships between these difficult concepts.

Geography 14 of the 15 participants mentioned the importance of geography, citing a need for: "[A] wide array of different geographical...representations" (P8). Others spoke of a "Regional distribution" (P1), which we have included here.

In particular, emphasis was placed on geographic markers of socioeconomic status such as the 'Global South', 'indigenous communities', and 'low-income countries'. One participant noted: "Immigration status is tied so closely to socioeconomic status" (P2), while another noted that "[Geography] is connected to socioeconomics because we know there are some poorer countries and rich countries" (P7). Others still asked questions like: "Do they have a passport?.... Are they in a refugee camp?" (P5).

Furthermore, participants saw it as important that their programmes had: "Global reach" (P7). They expressed a desire for: "[A] diversity of people coming from a variety of places" (P7).

[22]Columbia's policy of socioeconomic stratification divides households into 6 strata; unlike traditional measurements such as income or quality of life, these strata only consider household location and accommodation and thus capture a different facet of socioeconomic status [32].

Race While 11 out of 15 participants identified race as an important dimension of diversity, none suggested they would explicitly select racial diversity. Several participants instead noted the difficulty of measuring race in a global context: "Racial categories obviously vary by country" (P10). One participant noted: "[In places] like Brazil or England, there are different categories of race than there are in the US...[In Brazil] there's a board of people who decide what people's race are" (P5). In a global context, however, many participants pointed to relationships between geography and race and hence suggested diversifying across geography in place of race: "If it's an international programme then you can use geography as a proxy" (P5).

Types of Thinking 4 out of 15 participants discussed a diversity in the types of thinking exhibited by applicants. One participant noted that: "You want as much representation from different types of thinking as you can, because I want perspectives to be listened to equally" (P2). This manifested in many ways.

Participants tended to express the belief that personality type diversity could improve group cohesion: "With that understanding [of] personality types...be able to tell which...people would get on well with each other" (P14). One participant suggested a "Personality test" (P12), and another specifically mentioned a desire to diversify across "Openness" (P2).

However, while personality type was seen as important, core beliefs were seen as even more so. One participant noted an interest in the diversity of "Interests politically" (P12), and expressed a desire for diversity of "People's core beliefs...separate from religion" (P12). Another also noted: "I would try to have a good representation of...religious groups" (P3).

6.4.2.3 Operational Risks and Considerations

While our study did not focus on the operational aspects of selection, several selectors' understanding of diversity was closely tied to the operational realities of selecting for and running a scholarship. In answering our questions, several participants identified operational risks or considerations that impacted their understanding of diversity. These are listed as subthemes of 'Operational Risks and Considerations' in Table 6.1.

Outreach While our study was focused primarily on selecting a diverse cohort from a fixed applicant pool, 6 out of 15 participants answered questions from the perspective of outreach to grow a more diverse pool of applicants to select from. In particular, participants suggested that "Using technology for....targeted outreach" (P4) could help improve overall cohort diversity before selection even begins. One said: "Giving you very clear signposting on where you may want to focus, you know further recruitment or outreach or whatever it might be to make sure that your programme is diverse at the end of the day" (P6). Another added: "You can target your outreach dollars to communities where you know that underrepresented talent exists" (P10).

Support Similarly, 8 out of 15 participants suggested that technology could enable the support of applicants from underrepresented groups, which would also improve diversity. One participant suggested that technology could be used to provide: "[Support] to keep people that you're attracting from underrepresented backgrounds and help them get across the finish line" (P10). Participants focused on the: "Support needed to actually get [applicants] through your programme" (P10), i.e., supporting applicants after acceptance. Another suggested that technology could be used to provide support to applicants "After selection" (P15).

Selectors 4 out of 15 participants noted that diversity did not apply merely to applicants. Instead, for programmes where a group of selectors assists in the selection process, "Tracking the diversity of the selectors" (P15) and "[Monitoring] how they're scoring and reviewing applicants [for] prejudice or biases" (P15).

Applicant fraud Finally, only 2 out of 15 participants expressed concern with selecting based on particular diversity characteristics, especially self-reported metrics of diversity characteristics, was the potential for applicant fraud, e.g. falsely reporting demographic or other attributes to increase their chances of acceptance. One participant requested: "A fraud detector" (P5), while another expressed a desire to ensure that the process "Isn't super gameable" (P10).

6.4.2.4 Fairness and Bias

Though not a type of diversity as we have understood it here, many participants referenced similarities between diversity metrics and metrics of fairness or bias (as in Zhao et al. [187]). Furthermore, several suggested that improving fairness while reducing bias would likely yield a more diverse cohort. These themes are reflected under 'Fairness and Bias' in Table 6.1.

Fairness 6 out of 15 participants discussed that it was important that applicants: "Get fair chance on their on their academic merit" (P7). This translated to an emphasis on "Fairness in the assessment" (P7).

However, participants also noted that "The way the world works is unfair" (P14), and found it important that the programme is: "Making sure that the world is fairer by bringing more diversity to this world" (P7). In this way, participants found: "[The] representative thing....goes back to fairness" (P15). One participant noted that: "Affirmative action....can come across as unfair to some people, but....it's trying to balance things out when things have been so unequal for so long" (P3). This supports Zhao et al. [187]'s positioning of fairness as related to, rather than in opposition to, diversity.

Bias 11 out of 15 participants discussed bias, often as both a human- and machine-decision-making problem. Many participants appealed to technology's ability to be comparatively impartial as an important mitigator of bias, i.e., one participant repeatedly requested a: "Non-biased programme" (P3); another stated a preference for: "Data analysis to make decisions on who we should be supporting as opposed to having humans try to make those decisions with all their biases" (P5). However, those same participants noted that common machine decision-making paradigms amplify bias and that it was important to be aware of this: "AI has a lot of bias in it" (P3).

Others noted the possibility for technology to elucidate biases in both humans and machines. One participant requested: "Some kind of tool that can detect bias in a selection" (P7).

6.4.2.5 Programme Goals

Selectors from both programmes identified goals for their programmes. These goals were discussed by both groups as an intended form of 'impact' and both groups closely related achieving their goals to their expressions of why diversity mattered. While impact goals varied based on the type of impact and the affected party, we discuss these as 'impact' in Table 6.1.

Impact 5 out of 15 participants found key goals of their programme to include: "Orient[ing] [scholars] towards social impact or using their talent for good" (P8). They tended to encourage: "[Scholars'] working towards something impactful throughout their career" (P8).

6.4.2.6 Merit

Several participants reflected on the relationship between merit and diversity. While some participants saw these as competing goals, others saw them as complementary. In particular, complementary views often viewed merit as a form of performance relative to specific advantages or noted that many of our measurement tools are biased across our chosen diversity dimensions. These themes are reflected as subthemes of 'Merit' in Table 6.1.

Performance relative to disadvantage 2 out of 15 participants identified merit as something difficult to disentangle from performance. One participant noted that applicants may appear less qualified because they: "Didn't have the chance; didn't have the opportunities" (P2), while others with the opportunities will appear more qualified. Another participant began by asking: "How good are their three A stars based on where they've come from?" (P12), then proceeded

to reflect that "Your performance relative to your opportunity or maybe expected performance" (P12) is a key indicator of merit.

Measurement Closely related, 3 out of 15 participants questioned our ability to measure merit independent of opportunity: "[Whether they perform well] because they have the opportunity or because they are brilliant – I think that these two are really difficult to untangle" (P7). Another noted that: "Contextual factors mess up our otherwise seemingly objective measures of merit.... national context and family income is messing up your ability to measure the thing you actually care about" (P10). They continued to note that they: "Need to pay attention to [diversity] because it's messing up your measures of what you actually care about" (P10).

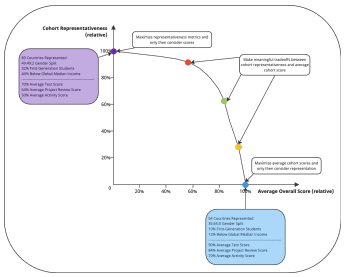
Performance Finally, 2 out of 15 participants noted occasions where performance and diversity were ostensibly competing goals. However, even here, participants recognised that observed performance and actual merit may differ. One participant noted that: "[The] overriding aim is for [the programme] to be as diverse as is possible but still meet a standard....relative score of like how good their application is based on all these kind of contextual factors" (P12). Another requested a technology that helps discover how: "Close you are to your idealised diversity targets and how close you are to maximising whatever it is you think you're maximising in your performance scores" (P10).

6.5 Study 2: Participatory Design

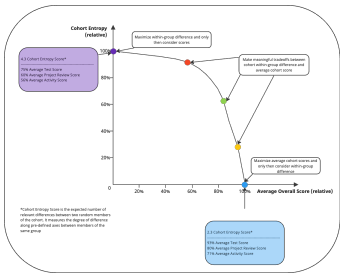
6.5.1 Prototypes

As a next step, we applied the results of our thematic analysis to design six prototypes following methodology from Buchenau and Suri [27]. These technologies aimed to help selectors better understand and operationalise diversity in their selection processes. We then present these prototypes to participants in participatory design workshops. These prototypes are shown in Figure 6.3.

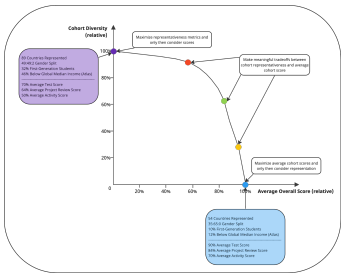
Three of these prototypes (Figures 6.3a, 6.3b, and 6.3c) present information about the range of possible cohorts participants must choose between. In addition to the themes uncovered in Section 6.4, these prototypes draw from the economic theory presented in Chapter 7. Meanwhile, the other three (Figures 6.3d, 6.3e, and 6.3f) present information about an individual applicant relative to a given cohort and a given pool. Furthermore, five of the six prototypes were designed to satisfy definitions of diversity uncovered in Section 6.4.



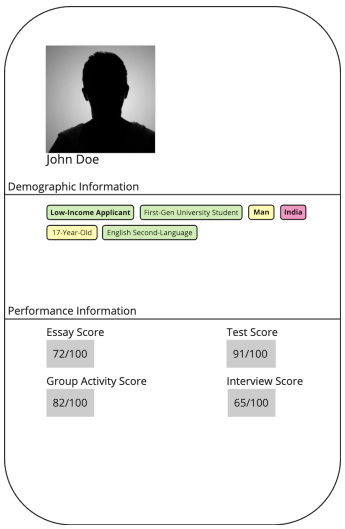
(a) Prototype 6.3a: Cohort Representativeness



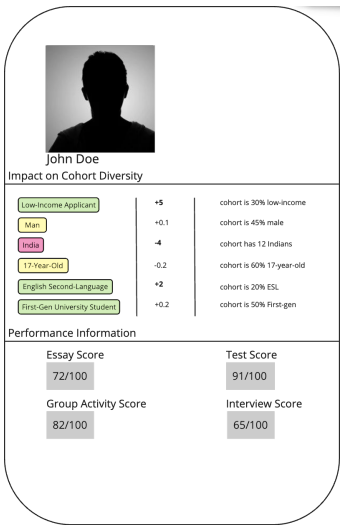
(b) Prototype 6.3b: Cohort Entropy



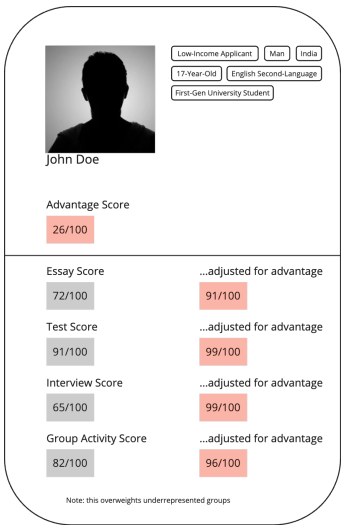
(c) Prototype 6.3c: Cohort Diversity



(d) Prototype 6.3d: Applicant Demographic Information



(e) Prototype 6.3e: Applicant Demographic Impact on Cohort



(f) Prototype 6.3f: Applicant Advantage Scores

Figure 6.3: These figures depict the prototypes designed based on themes from Section 6.4 and used in our participatory design workshops. They are reproduced at a larger scale in Appendix D.2

Figures 6.3d and 6.3e are based on the 'representativeness' theme, while Figure 6.3f is based on the 'contextualising applications' theme. Similarly, Figures 6.3a and 6.3b draw a distinction in their measurements of diversity, based on the

'representativeness' and the 'different perspectives' themes, respectively. Figure 6.3b, in particular, defines and employs 'entropy' as a metric. This reflects that when aiming to get different perspectives in the same room, the goal is not to represent any target population; rather, we desire that everyone in our group be as different from the remainder of the group as possible.

6.5.2 Workshop Methodology

As the participants come from two separate talent investment programmes, we run one workshop for each group [27]. Before presenting to the broad audience of each group, we submit our figures to one primary contact (also a participant in the study) at each organisation, then run informal, 15-minute 'pilot' one-on-one workshops with this primary contact. Here, we primarily sought approval to use these figures in a workshop with the broader team of selectors, but we also collected minor feedback and tweaked the prototypes based on this feedback. In one organisation's pilot workshop, the primary contact requested Figures 6.3a and 6.3b be combined into one prototype with 'Diversity' as the Y-axis, as the organisation already has an internal working definition of diversity (that incorporates what we mean by both representativeness and contextualising applications). This can be seen in Figure 6.3c. Participant grouping for the workshops is redacted to preserve participant anonymity; the results for this analysis are attributed by group, rather than individual, to reflect the cooperative nature of the task.

Our central research questions for this workshop are:

1. What prototypes best promote diversity?

2. What elements of these prototypes facilitate their success?

Or, for each prototype: "How and why does this prototype promote diversity?". In each workshop, we ask participants to consider each prototype in turn and to discuss how they might use it in their selection process. We then ask participants to consider how these prototypes might fit into their current selection process, and how they might change their process to better incorporate these prototypes. Finally, we ask

participants to consider how their current selection process might make the best use of these prototypes, and whether they think these prototypes would be beneficial.

Following the methodology of Gatian [56] and Griffiths et al. [61], at the end of each workshop, we ask participants to highlight their favourite prototype.

A question-by-question protocol for these workshops can be found in Appendix B.3.

6.5.3 Results

6.5.3.1 Participants Preferred Prototype 6.3e

As part of the workshop, participants were asked to mark their favourite prototypes [56, 61]. These favourites have been collated in Table 6.2, and it can be seen here that Prototype 6.3e was by far the favourite in both groups.

Table 6.2: This table tallies the number of participants who indicated that a given prototype was their favourite. The overwhelming favourite was prototype 6.3e, which shows an individual applicant's impact on the cohort.

Prototype	Favourites
Prototype 6.3a	1
Prototype 6.3b	1
Prototype 6.3c	0
Prototype 6.3d	1
Prototype 6.3e	10
Prototype 6.3f	0

6.5.3.2 Both Groups Rely on Idiosyncratic Notions in Selection

When participants were placed in a practical selection scenario and shown technology prototype information about (hypothetical) applicants, they compared applicants to the idiosyncratic profiles that they desired.

When evaluating Prototype 6.3f, G1 used advantage scores to seek out: "Diamonds in the rough" (G1), i.e., talented applicants from disadvantaged backgrounds who lacked the polish of their more privileged counterparts.

Prototype 6.3b was initially confusing to G2, as the 'entropy' definition used was unfamiliar to the participants: "Entropy is chaos in chemistry. How does this

relate to our usage here?" (G2). Thus, when interacting with Prototype 6.3b, G2 understood 'entropy' to be a variety in cognitive skill and personality type. In particular, the group sought to ensure that the chosen cohort contained 'glue' people, who improve overall cohort cohesion: "For people working together, it's useful to have someone who is that 'glue"' (G2).

6.5.3.3 Organisations and Participants Are Interested in Different Diversities

When presented with Prototypes 6.3a and 6.3b, G2 was given the demonstrative definitions of 'representativeness' and 'entropy' (visible on the prototypes). G2 quickly understood the differences between the figures to be that Prototype 6.3a sought to support considerations of demographic representativeness, while Prototype 6.3b sought to support placing different perspectives (be they cognitive skill-sets or personality types) in the same room. Participants were then interested to know the relationship between these prototypes in practice: "If we maximise based on [Prototype 6.3a] scores, what would the Entropy scores be?" (G2).

However, though individual participants took great interest in Prototype 6.3a, G2 ultimately acknowledged that the programme was most interested in building a cohort from people with different perspectives to facilitate collaboration: "For [our] cohorts, [we] want a balance [of personality types] and [we] want them to be collaborative" (G2). At the same time: "Let's track but not use [programme-specific measure of representativeness]" (G2).

6.5.3.4 Different Tools are Useful at Different Application Stages

Participants in both groups variously expressed anxiety about measuring the relevant dimensions. "What are our metrics and are they reliable?" (G2) was echoed by several G2 participants. "If it's all self-report, then we can't do anything with it" (G1).

However, both organisations noted that their selection processes involve a variety of stages. Different tools are useful in different stages. When speaking of Prototype 6.3b, one participant said: "This is better post-interview than it is pre-interview"

(G2), as interviews will collect observational data on many of these characteristics. I.e., while certain tools may rely on measurements that cannot be collected until later stages, others will be more useful earlier in the selection process.

Similarly, different output modes were useful at different stages. Note that Prototypes 6.3d and 6.3e contain similar information, but that Prototype 6.3e displays this information in greater detail. When reviewing both prototypes, G1 found Prototype 6.3d preferable in the earlier stages of decision-making, while Prototype 6.3e had the greatest utility later. "[We] can't send [Prototype 6.3e] as a pre-read, but [Prototype 6.3d] makes more sense in isolation, so better for a pre-read" (G1). On Prototype 6.3d in particular, one participant noted: "Helpful for the process, not so much for the [final cohort selection]" (G1). Another said of Prototype 6.3e: "This has the most potential at the later stages of decision-making" (G1).

At the other extreme, while the cohort-level tools did not spark discussions about particular individuals, they did spark earlier-stage, higher-level discussions. Most obviously, both programmes discussed specific tradeoffs between measured individual applicant aptitude and cohort diversity: "Real decision-making always sees tradeoffs like these.... This chart helps you figure out the level of compromise you're willing to make on both axes" (G2); "[It] makes sense that top scoring candidates don't necessarily help you build the most diverse cohort.... 5-10% [of the cohort] really get to a struggle between quality and diversity.... If this chart were real (rather than hypothetical), and you could see who you were losing, this would be useful" (G1). (In the hypothetical, participants from G1 settled closer to the centre of the frontier: "[Our] target here is to look somewhere [from] red to yellow" (G1). However, they also noted that "Some candidates get a big diversity boost and score terribly" (G1).)

In one case, participants also discussed broader, programme-level concerns. Participants spent time debating "Is the programme needs-based or merit-based?" (G1). They noted that "This chart helps [facilitate that discussion]" (G1) but did not ultimately conclude one way or the other.

6.5.3.5 The Right Balance of Quantitative and Qualitative Information is Key

Participants simultaneously expressed gratitude that the prototypes were as simplified as they were, and a desire for more detail. One participant from G1 noted that the prototypes are: "Very constrained in terms of what is being shown.... This doesn't include all of the factors, but for the decision, that's good, because it prevents info overload" (G1). However, participants G2 frequently requested additional quantitative information. In the case of the Prototype 6.3e (participants' favourite prototype, as can be seen in Table 6.2), participants requested many other metrics: "Advantage score" (G2), "Impact on entropy" (G2), "[A summary of] these scores together" (G2), and "A composite impact on cohort diversity" (G2).

Meanwhile, participants from G1 requested qualitative information: "We need to know more about the applicants' backgrounds" (G1). Other requested information included: "Comments from selectors" (G1) and "A narrative summary...written by the selector team" (G1).

This suggests a discrepancy between the two groups' preference for the balance between qualitative and quantitative information. While G2 expressed a desire for unifying quantitative metrics: "A single score for disadvantage [or] need, while recognising its flaws, could provide one read of an individual's circumstances" (G2), G1 expressed a desire for individual, qualitative information: "We're using quantitative to sift through qualitative.... [We] need to include comments from selectors" (G1).

6.6 Design Recommendations

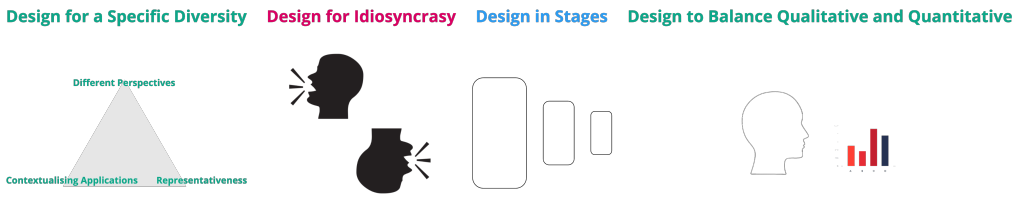


Figure 6.4: This figure illustrates our four key design recommendations to others building tools to support the selection of diverse talent.

6.6.1 Design for a Specific Diversity

In the initial interviews, participants were often vague when asked to define diversity. However, when asked to expand on why diversity is important, or on what dimensions of diversity they prioritised, it became clear that 'diversity' included three separate (and sometimes competing) definitions. We have termed these: 'representativeness', 'different perspectives', and 'contextualising applications'. When designing tools to assist a target organisation in considering diversity, we suggest designers first clarify through human-centric methods which definitions of diversity the target organisation seeks to consider, then designs to support those specific definitions. That is, designers should follow Van Kleek et al. [172]'s paradigm of 'respectful' design and build technology to best serve the needs of the selectors who will use it.

6.6.2 Design for Idiosyncrasy

One key note revealed through this process is that different decision-making processes have philosophical underpinnings, desiderata, and anecdotal definitions that impact their selections. These idiosyncrasies should be discovered early in development and designed for in any technical solution. We suggest participatory design as a mechanism for achieving this. We observed a strong relationship between participant feedback in interviews and their satisfaction with the prototypes.

For example, both talent investment programmes we worked with have created specific personas they look for. For example, one group discussed 'glue' people who helped groups function cohesively. The other group discussed 'diamonds in the rough', talented youth systemically undervalued due to their backgrounds. Where these aspects were included, participants showed strong interest in the prototypes. Where they were excluded, participants often asked for these to be added.

6.6.3 Design in Stages

Much of what participants desired at one stage of decision-making was mutually exclusive with what they desired at other stages. For example, G1 desired Prototype 6.3d as a: "Pre-read" (G1) due to its simplicity but preferred the detail of Prototype

6.3e “In the room” (G1). Thus, it is crucial for designers to consider what stage of decision-making their tools are designed to support and to design appropriate levels of detail, abstraction, and engagement accordingly.

6.6.4 Design to Balance Qualitative and Quantitative

Participants often noted that the prototypes were missing key qualitative information about applicants. This qualitative information is crucial in the holistic considerations of each applicant. However, when allowed to consider only qualitative information, participants obscure tradeoffs they are forced to make between different programme goals. In particular, while individual-level goals are often clear, cohort-level goals (such as diversity) are easier to delay or ignore. Thus, without quantitative tools to frame the discussion, participants noted that they were often forced to make cohort-level considerations ad-hoc and towards the end of their decision-making process.

Thus, while qualitative information is crucial, it is also important to present the quantitative information necessary to make these tradeoffs salient. Ultimately, final selection decisions are made by panels of trained selectors, but in the absence of both quantitative and qualitative information to guide these decision-makers, they would be forced to make decisions that are less well-informed than they could be.

6.7 Discussion

6.7.1 The Diversity Triangle

In Section 6.4, participants variously identified the word ‘diversity’ with three themes, which we have taken to signify definitions of diversity: ‘representativeness’, ‘different perspectives’, and ‘contextualising applications’. As these three definitions are central to our research questions, we privileged them over Section 6.4’s other themes in Section 6.5, where we engaged in scenario speed dating and experience prototyping designed to satisfy different definitions of diversity. In this section, we map the themes to the three definitions in the Diversity Triangle (Figure 6.5).

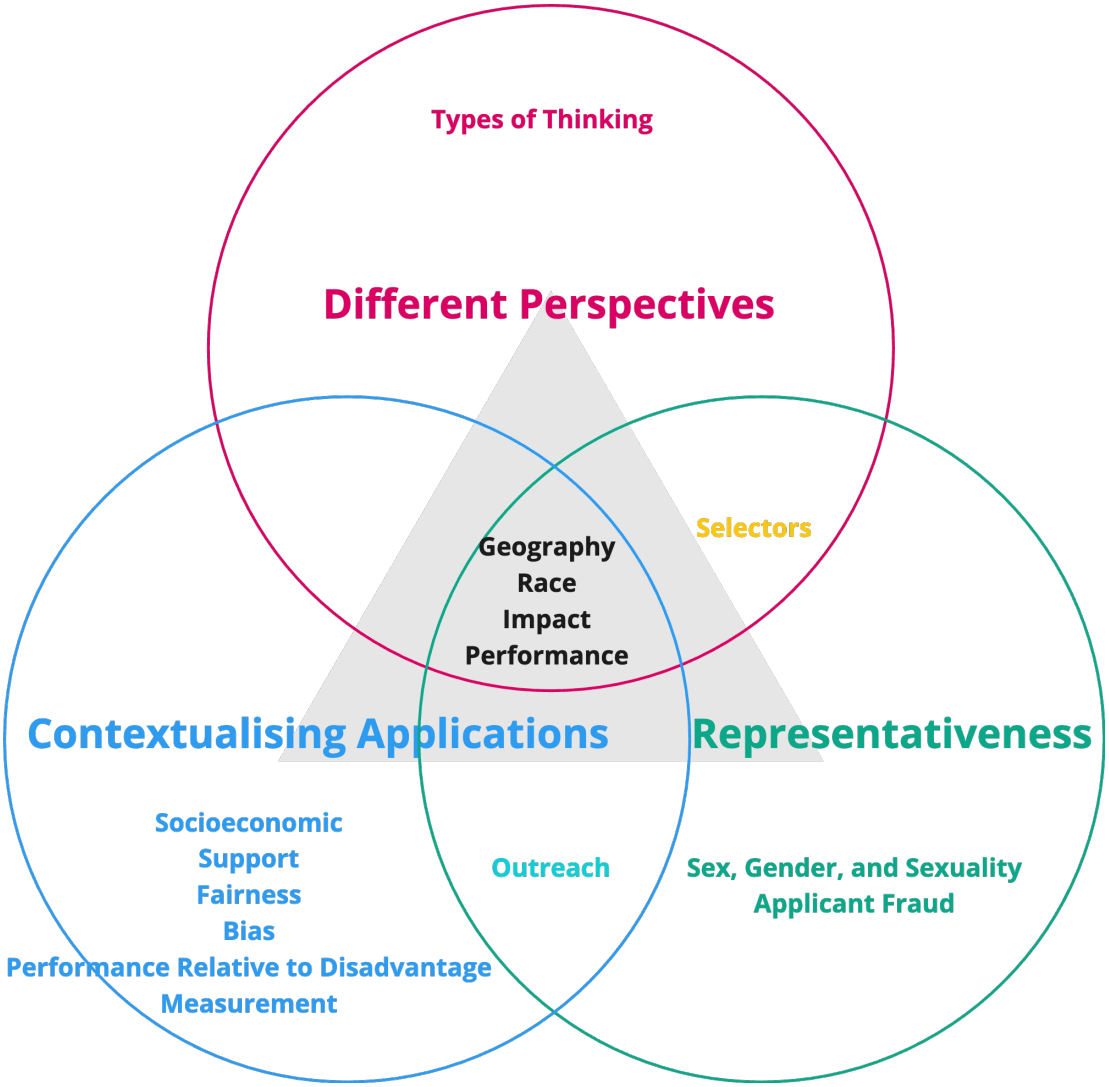


Figure 6.5: This figure depicts the Diversity Triangle, three differing definitions of diversity selectors expressed when discussing a diverse cohort. We also relate the Diversity Triangle to each other theme or subtheme participants mentioned.

6.7.2 The Relationship Between the Diversity Triangle and Theories of Change

6.7.2.1 Representativeness

Each of the definitions from the Diversity Triangle (Figure 6.5) implies different programme values and a different theory of change. The representativeness definition relies on social theories by which diversity is intrinsically valuable. Several participants seemed to believe representativeness here was intrinsically valuable, which aligns with the Morris [119] account under which contemporary diversity

norms emerged from loci of oppression affecting underrepresented groups. However, some participants also noted instrumental reasons to value representativeness that align closely with Page [132] and Peters [141]'s argument that people from different backgrounds often possess unique and germane knowledge. One participant noted that a team can: "Better serve a community if they represent [that community]" (P9). Another theory from Friedler et al. [52] discusses measurement bias. Notably, if we assume that talent is equally distributed across some partition, then the most talented cohort should also be representative. However, Friedler et al. [52] note that we often observe in practice that performance is not equally distributed across these partitions. This is likely due to both relevant differences between those groups and structural bias that causes differences in construct observability between groups (in fact, the relevance of said differences may also be due to structural biases) [52]. In this case, representativeness would be a proxy for the distributive notion of fairness [129]. Thus, representativeness is pro-social: society is better served when resources are distributed to a representative group of people.

6.7.2.2 Different Perspectives

The argument for placing different perspectives in the same room is often instrumental. While 'diversity' on the whole is often spoken of as an intrinsically valuable broader benefit, the aim of placing different people in the same room is often only to benefit the people in that room. Page [133] argues that 'cognitively' diverse groups outperform homogeneous groups on some tasks; some participants similarly contend that it improves cohort-level task performance. Other participants echo Wylie [183]'s argument that it allows participants to better learn from each other. In either case, the benefit is primarily organisational rather than social.

6.7.2.3 Contextualising Applications

The argument for contextualising applications is twofold. Most often, participants make a systemic critique here. That is, the world is incredibly unjust, and we want to distribute resources differently, but in a talent selection process, we still have to operate in an unequal world. Thus, to correct that injustice, we must give

more resources to those who have less. This could be seen as a form of distributive justice or 'affirmative action' [129]. Participants argue (perhaps relatedly) that appropriately contextualising applications results in more successful applications from disadvantaged groups; this, in turn, allows these applicants to have a positive impact on their groups. Relatedly, either due to measurement bias, or due to differences in performance brought on by disparate access to resources, support, and opportunities, we may find that applicants from disadvantaged backgrounds appear worse on paper. In either case, correcting this through contextualisation may also build a more fair selection process and a more just world, yielding great benefit to society. On the organisational side, if contextualising applicants allows for the admission of more applicants from marginalised groups, they may be better suited to critiques of existing power structures, as they may have more informative experiences of oppression [112].

6.7.3 A Cautionary Note On Designing for Diversity

Decisions such as scholarship selection have long-reaching impacts on the applicants. A successful applicant to a scholarship programme might thus attend a university they could not have otherwise attended; there, they will acquire skills and a network that will continue to impact them later in life; these impacts may even affect those close to them, as their better circumstances likely better the circumstances of their communities [39]. Thus, it is simultaneously important to ensure that these opportunities are dispersed fairly and to ensure that all demographics are included among those selected.

This increases the importance of selecting a diverse group of people and mandates that we do all we can to do so. Quantitative decision support tools may have a role to play in this, but two problems remain: data collection and data processing. Primarily, diversity data is often self-reported. Thus, we cannot use this naively to generate diverse cohorts, as doing so would yield a competitive advantage to candidates who lie on their declarations. Secondly, processing data automatically

has its drawbacks. AI systems are prone to bias [52]. And as these systems improve, it is unclear if they will help or harm our ability to select diverse cohorts.

However, it is important to note that both the problems of data collection and processing are not unique to this workflow. Algorithmically-supported diversity considerations may correct for data collection and processing issues elsewhere. Bias is a major consideration in data collection [52]; heterogeneous biases in measurements of talent may be corrected by ensuring diversity across these metrics: "Assessments are usually biased measures of what we care about, and that opportunity often correlates with positive error terms in assessments' measure of underlying skills.... [Diversity considerations] correct for the inadvertent affirmative action against underprivileged individuals implicit in using biased assessments" (P1). Similarly, much like AI systems, human selectors are prone to their own heterogeneous biases; in some sense, 'design for idiosyncrasies' reflects the notion that decision support systems must at times help recognise and mitigate these biases.

6.8 Limitations and Future Work

In Studies 1 and 2 we build tools with which participants report satisfaction, but increasing participant satisfaction does not necessarily improve decision-making. In particular, technology that makes difficult decisions less painful may be well-received, while technology that makes these decisions more salient may be less popular but more impactful [100, 111]. Our themes speak to participant perceptions of diversity and our design recommendations speak to participant desiderata from support tools. We assume that, in solving for these considerations, we can help organisations select better cohorts. However, we may find that these tools fail to improve decision-making concerning diversity in practice. Future work should investigate this possibility through the implementation of our prototypes in field settings.

Venn-Wycherley et al. [174] contend that HCI literature focused on educational contexts should consider both educator and student. Two distinctions distance our work from theirs: first, selection is distinct from pedagogy in that our decision subjects are not necessarily beneficiaries (while students are beneficiaries of their

institution, applicants do not become beneficiaries unless they are selected); second, scholarships are distinct from educational institutions in that scholarship programme benefits primarily focus on assisting beneficiaries in accessing educational institutions, while educational institutions primarily educate beneficiaries. Both distinctions create distance between selectors and applicants beyond that between teachers and students and pose challenges in engaging decision subjects. Nonetheless, future work should engage scholarship applicants to understand their definitions, stances, and considerations concerning diversity.

While not a limitation, we have intentionally set aside themes such as 'outreach' and 'selectors' that relate more to other aspects of selection processes than to the act of selection itself. We hope future work will consider these facets of selection programmes, especially when designing tools to support thinking around diversity.

6.9 Conclusion

This research answers the crucial question: "What is diversity?", from the perspective of scholarship and talent investment selectors. In doing so, we illuminate the multifaceted nature of diversity in scholarship selection and emphasise the critical need for tools that support considerations of diversity in decision-making. Our findings reveal that achieving true diversity involves navigating the complex interplay between three occasionally conflicting definitions: representativeness, diverse perspectives, and the contextualisation of applicants' backgrounds. By engaging in participatory design, we build six prototypes with our selectors; this process reveals four design recommendations: design for a specific diversity, design for idiosyncrasy, design in stages, and design to balance quantitative and qualitative. This work demonstrates that, when thoughtfully designed, technology can empower selection processes to be more equitable, inclusive, and transparent. The broader implication is that such advancements have the potential to reshape how diversity is operationalised, ensuring that it is both a measurable outcome and a core value in shaping the future of talent identification. As we move forward, the integration of these tools into real-world practices will be pivotal in fostering truly diverse

and representative groups in global scholarship programmes and beyond. However, we caution that programmes seeking to implement these tools should ensure that they improve not only subjective selector perceptions of decisions but also the objective quality of those decisions themselves.

7

A Possibility Frontier Approach to Diverse Talent Selection[23]

Contents

7.1 Motivation

In Chapter 6, we co-designed six prototypes with selectors from Programme A and Programme B, all of which garnered interest. However, as we caution in

[23]This chapter is based on two papers showcasing research done in concert with Kadeem Noray. Both contributed equally to the research. This version draws from both publications; in doing so, it borrows from the rich tradition of talent-related research in economics, as well as the HCI work referenced in Chapters 4, 5, and 6.

Chapter 2 and demonstrate in Chapter 4, subjective feedback is not a substitute for objective evaluation of a tool's impact on decision-making. Therefore, this chapter moves from design to deployment. We implement a functional version of Prototype 6.3c from the previous chapter and evaluate its real-world impact in a field deployment with the Programme A programme.[24]

7.2 Introduction

The rise of diversity, equity, and inclusion initiatives suggests that various organisations (e.g., schools, firms, social impact programmes, etc.) are genuinely interested in selecting diverse talent. This is driven, at least in part, by recent declines in discrimination [70], increases in the perceived return to diversity [42, 128, 134], and increased social pressure for demographic representation [113]. In this chapter, we deploy technology based on Prototype 6.3c in a field study with Programme A. We seek to understand whether this technology can help organisations select more diverse and talented cohorts.[25]

To ground and assess our technology, we first develop a model for cohort selection. We assume an organisation receives N applications and must select a cohort of $n < N$ individuals. The organisation seeks to simultaneously maximise the cohort's talent (e.g., mean performance on an ability measure) and its diversity (e.g., proximity to target demographic proportions). This optimisation problem yields a trade-off, which we term the Selection Possibilities Frontier (SPF). The SPF represents the set of non-dominated cohorts—those that are not outperformed by any other cohort on both talent and diversity. The SPF framework provides the

[24]It should be noted here that Prototype 6.3c itself draws from theory presented in Section 7.3.1. Thus, the design of Prototype 6.3c implemented in this chapter draws as much from this chapter as from Chapter 6.

[25]Throughout the chapter, we use 'talent', 'aptitude', and 'performance' interchangeably. In doing this, we recognise that organisations generally seek to optimise for vague and often conflicting notions of talent or aptitude, but do so via measurements of applicant performance on assessments or assignments. In practice, this process carries risks; e.g., heterogenous biases in metrics or assessment methods will lead some subgroups to appear less talented or apt than others, even when no true difference in talent or aptitude exists. On the whole, this chapter is more interested in measurements of diversity than of talent or aptitude; thus, while these risks are of crucial importance to fair selection processes, they are tangential to the focus of this chapter.

theoretical basis for Prototype 6.3c, which we implement using a greedy estimation algorithm that leverages the submodularity of diversity functions [71, 86].

Next, we present the results of deploying this tool with Programme A during their Cycle Y selection cycle. Our analysis of their Cycles W and X reveals that their selected finalist cohorts were well within the SPF, meaning they could have been substantially more diverse or higher-performing. For instance, the Cycle X cohort could have been 14.6% more diverse with no loss in performance, or 24.1% higher-performing with no loss in diversity. This indicates that without our DST, Programme A was making inefficient selections relative to their own stated goals. In contrast, when Programme A used the SPF-based tool in Cycle Y, they selected a cohort that was more diverse, higher-performing, and located very near the estimated frontier. This suggests the DST significantly improved the efficiency of their selection process.

To understand why organisations might select inefficient cohorts without such support, we prove that the underlying optimisation problem is computationally complex. When diversity preferences involve non-mutually exclusive identities (e.g., both ethnic minorities and women), the problem of finding the optimal cohort becomes **NP**-hard. When these preferences are considered among an organisations myriad other preferences, the problem remains **NP**-hard. This complexity makes it prohibitively costly for selectors to find the optimal frontier by hand, leading them to choose suboptimal cohorts. Our DST reduces this computational cost, enabling them to make more efficient decisions.

To account for this complexity, we augment our model of cohort selection by forcing organisations to incur a computational marginal cost for each unit of increased cohort diversity. This represents the fact that, to find a more diverse cohort, one must engage in the laborious process of composing potential cohorts and comparing them. This is in sharp contrast to finding more talented cohorts, which is comparatively simple and thus modelled as costless, because each individual's contribution to cohort talent is unrelated to the remainder of the cohort. This updated model appears to better describe organisational behaviour. This has two

implications: (1) organisations will tend to select sub-optimal (i.e. non-first-best) cohorts and (2) organisations will improve on both performance and diversity if they gain access to a technology that reduces this computational marginal cost.

As an aside, we apply our SPF estimation procedure as an ex-post DST to evaluate the efficacy of alternative screening and selection methods. To do this, we leverage two unique aspects of the programme. First, the programme collects both traditional merit-based measures – including cognitive tests, written essays, and referring organisations – as well as non-traditional measures – including peer-reviewed video essays, gamified skill tests, and application platform behaviours. Second, the programme engaged in effectively no screening before receiving concrete projects from applicants, making it possible to estimate valid counterfactual diversity and performance of cohorts had they been screened in different ways. Leveraging these features, we find three key results. First, selecting only based on cognitive ability or traditional metrics would have improved cohort performance relative to random selection, but would significantly restrict the programme from reaching its diversity goals. By contrast, selecting only on peer reviews performs similarly on performance but improves diversity substantially. Second, all alternative selection methods we explore result in selecting cohorts well within the SPF and, therefore, leave substantial diversity and performance gains on the table. Third, the trade-off implied by the SPF between talent and diversity is steeper if traditional measures are used to measure talent than if applicant projects are used.

7.3 Theory and Methods

7.3.1 Evaluating Organisational Decision-Making

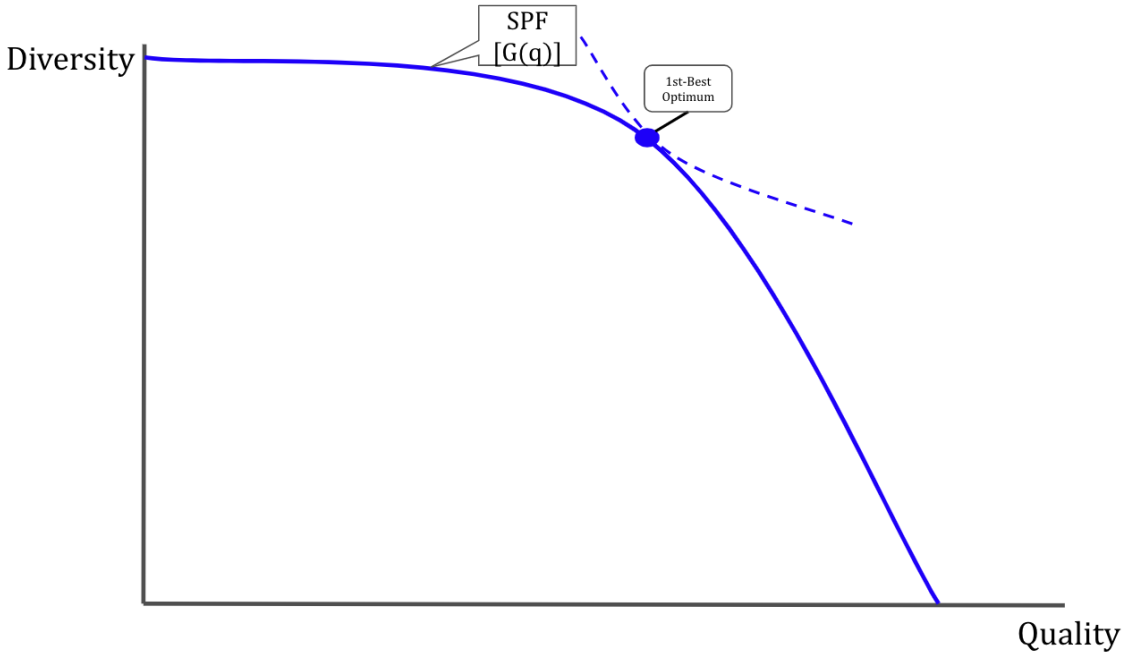
Central to this chapter is the desire to evaluate Prototype 6.3c on real results, rather than subjective satisfaction. However, there is no ground truth of decision-making in the scholarship programme. Thus, we define a model of talent selection and use this model in assessing a field deployment of the prototype. This model relies on a key object, the Selection Possibilities Frontier (SPF), to bound the range of possible cohort selections on two axes: average applicant performance on individualised

metrics of talent, and overall group diversity. With this model, we can evaluate the programme's decisions in terms of their proximity to the SPF; the closer a programme is to the SPF, the more efficient its selection process. In this case, we deploy Prototype 6.3c with Programme A, thus, we are interested in Programme A's SPF. Programme A does not share the precise aggregation method for their talent and diversity metrics, but a summary of the measurements they collect can be found in Appendix A. For this chapter, it suffices to know that Programme A has working definitions of both performance and diversity.

We start by considering a simple version of the organisation's optimisation problem. Organisations receive N applications and must select $n < N$ individuals to form a cohort c from the set of all potential cohorts C. The organisation prefers both that the selected cohort is higher-performing on some measure of talent and more diverse. For now, a cohort's diversity can be thought of as the inverse of a multidimensional measure of distance between the set of proportions of the cohort who belong to key demographic groups and a set of target proportions the organisation has for each group (we discuss definitions of diversity in more detail in Section 7.5.1). If we let the performance and diversity of a given cohort c be given by the functions $P(c)$ and $D(c)$, respectively, then the above description is equivalent to letting the organisation's preference function $F\Big(D(c), P(c)\Big)$ exhibit $F_D > 0$, $F_P > 0$, and $F_{DP} \geq 0$, where subscripts indicate partial derivatives.

Conceptually, this would represent a scenario where an organisation can observe D and P for every possible cohort and simply select the one that maximises $F(D, P)$. If we assume the organisation behaves rationally, we know the organisation will not choose dominated cohorts. Formally, c^* can be the optimal cohort if and only if there exists no c' such that $D(c') > D(c^*)$ and $P(c') \geq P(c^*)$ and there exists no c' such that $D(c') \geq D(c^*)$ and $P(c') > P(c^*)$. We know that the optimal cohort must be in the set of non-dominated cohorts which we define as the Selection Possibilities Frontier (SPF). If we assume, for expositional purposes, that the SPF is continuous, we can represent it as the following function:

Figure 7.1: This figure depicts an example solution to an iteration of the selection problem, which is described in Equation 7.2. The solid blue curve represents the Selection Possibilities Frontier (SPF), the dotted blue curve represents the organisation's indifference curve corresponding to the highest achievable utility, and the blue dot represents the diversity and performance of the optimal choice (i.e. the first-best solution).



$$G(p) := \max \Big[D(c) | P(c) \geq p \Big] \tag{7.1}$$

In words, $G(p)$ merely gives the highest possible diversity for every cohort performance level. The diverse talent selection problem, then, can be represented as choosing a cohort to maximise F subject to a constraint that the choice be on the SPF. Formally:

$$\max_{d,p} F\Big(d, p\Big) \quad \textbf{s.t.} \quad d = G(p),$$

which is equivalent to:

$$\max_{p} F\Big(G(p), p\Big). \tag{7.2}$$

The solution to this simple version of the model is depicted in Figure 7.1. Notably, this model suggests that organisations should always select cohorts on the

frontier, as all cohorts within the frontier are dominated by cohorts that are either at least as diverse and more talented or at least as talented and more diverse.

7.3.2 Defining the Classes of Diversity and Performance Functions

Without knowledge of the types of functions $D(c)$ and $P(c)$, Equation 7.2 proves difficult to instrument in practice. In this section, we define the classes of functions $D(c)$ and $P(c)$ that are relevant to the diverse talent selection problem. In particular, we will formalise *proportional diversity* and *count diversity* as kinds of diversity, and assume performance to be a real-valued individual-level metric, aggregated by summation.

In both cases, here, we work with Programme A to identify their preferences w.r.t. different kinds of diversity and performance. We then implement their preferences as functions $D(c)$ and $P(c)$.

Proportional diversity When organisations make statements like: "we desire at least x proportion of group g", they are speaking of proportional diversity. But, since organisations aim to select cohorts of a specific size, we can reframe this goal as "we desire at least $x * n$ individuals from group g", where n is the total number of applicants in the cohort.[26] If we let $\chi_g(c)$ be the proportion of c in group g and $\sigma_g(c)$ be the total number of applicants in c who are in group g, this goal can be formalised into the proportional diversity function:

$$\begin{aligned} \delta_g^{prop}(c, x) &:= n * \min(\chi_g(c), x) \\ &:= \frac{n * \min(\sigma_g(c), x * n)}{n} \\ &:= \min(\sigma_g(c), x * n). \end{aligned} \tag{7.3}$$

Note that the minimum function is used here to formalise "at least", so that the function only increases until the proportional threshold is met. If, for example, an organisation selecting 100 applicants would like their organisation to be at

[26] This reframing will turn out to be helpful in Section 7.3.3 when we develop our SPF estimation strategy

least 40% female, the proportional diversity function associated with this goal is $\delta^{p}_{female}(c, 40) := \min(\vec{\mathbf{f}} * \vec{\mathbf{c}}, 40)$ where $\vec{\mathbf{f}}$ is a Boolean vector indicating which applicants are female and $\vec{\mathbf{c}}$ is a Boolean vector that indicates who is in cohort c. This can be thought of as an inverse distance along the dimension of group representation between cohort c and an ideal cohort c^* where $\sigma_g(c^*) = x * n$.

Count Diversity The second common type of diversity is what we call *Count diversity*. This formalises organisational statements like: "We desire at least one person from m groups". To formalise this notion, let $\mathbb{I}(\cdot)$ be an indicator function that is equal to 1 if the condition within is true. We can now represent a count diversity preference as:

$$\delta^{count}_{G}(c, m) := \min\Big(\sum_{g \in G} \mathbb{I}(\sigma_g(c) \geq 1), m\Big), \tag{7.4}$$

where G is the set of relevant groups the organisation wants to be represented by at least a single individual. This type of function is ideal for representing geographic representation goals where educational institutions, like colleges or scholarships, often have goals like "we want a student from every state" or "we want as many countries as possible represented".

Overall Diversity Ultimately, organisations care about all of their diversity goals, not just one. Thus, the diversity functions that are relevant for an organisation must be aggregated if we want to formalise an organisation's overall preference for diversity. We define this aggregation as an organisation's diversity score $D(c)$, which generally has the following form:

$$A\Big(\delta^{prop}_{g_1}(c, x_1), ..., \delta^{prop}_{g_K}(c, x_K), \delta^{count}_{G_1}(c, m_1), ..., \delta^{count}_{G_J}(c, m_J)\Big),$$

where $A(\cdot)$ is an aggregator function. It is essential that $D(c)$ increases as a cohort gets "closer" to one of the underlying diversity goals because this is sufficient to identify cases when one cohort dominates another, even if the formalisation

misses something subtle or difficult to articulate about the organisation's diversity preferences. A flexible but simple aggregator function is a weighted sum, where organisations can place different emphases on each of the goals. So, for the remainder of this chapter, we use diversity scores of the following form:

$$D(c, \vec{\mathbf{w}}, \vec{\mathbf{x}}, \vec{\mathbf{m}}, \vec{\mathbf{g}}, \vec{\mathbf{G}}) := \sum_{k \in K} w_k \delta_{g_k}^{prop}(c, x_k) + \sum_{j \in J} w_j \delta_{G_j}^{count}(c, m_j), \tag{7.5}$$

where $\vec{\mathbf{w}}, \vec{\mathbf{x}}, \vec{\mathbf{m}}, \vec{\mathbf{g}}, \vec{\mathbf{G}}$ are vectors of the organisation's weights, proportional targets, count targets, groups of interest to proportional diversity functions, and sets of groups of interest to count diversity functions, respectively.[27] In general, we suppress the vector notation opting to refer to the diversity score as $D(c)$ where this doesn't lead to confusion.

Talent, Aptitude, or Performance Relative to diversity, our definition of performance is simple. In general, organisations measure aptitude for their programme using an individualised metric, usually performance on some assessment or assignment. Common examples include test scores, essays, or grades for educational organisations or technical interviews for hiring in technology. More sophisticated (though uncommon) measures might be the predicted success of an individual based on a set of performance metrics. In this chapter, we assume that organisations already possess a real-valued talent metric ρ_i evaluated at an individual level. A cohort's *talent*, then is defined as the sum of the talent level of the individual members, which is given by:

$$P(c) := \sum_{i \in I_c} \rho_i, \tag{7.6}$$

where I_c is the set of all individuals i in cohort c. Unlike diversity, $P(c)$ is straightforward because each individual's contribution is ρ_i regardless of whoever

[27] Another attractive option is a CES aggregator because it allows for specifying the degree of substitutability between diversity goals, but this comes at a cost to interpretability, as many organisations don't regularly use CES aggregators. Nonetheless, the authors are currently working on establishing whether the estimation procedure presented in Section 7.3.3 is viable for a CES aggregator.

else is in the cohort. Note that, as long as an organisation fixes their desired cohort size beforehand, optimising for the sum of ρ_i is identical to optimising for mean ρ_i.

We represent an organisation's preference function F as a weighted sum of performance and diversity functions. That is:

$$F(D, P, c, \iota) := \iota * D(c) + (1 - \iota) * P(c) \tag{7.7}$$

7.3.3 Implementing Prototype 6.3c in the Field

As it happens, the SPF modelled in Figure 7.1 can be used to implement Prototype 6.3c in the field. I.e., a calculation of the SPF using an organisation's preference function yields all of the data required to plot Prototype 6.3c (which is, as it happens, just a depiction of the SPF presented alongside contextual information designed to help selectors best understand the visualisation). However, as we will demonstrate in Theorem 1, calculating the SPF outright is unfeasible.[28] Instead, we rely on a greedy algorithm to approximate the SPF.

Greedy optimisation is the practice of approximating an optimal solution to an iterative process by, at each iteration, making a choice that optimises the process at that iteration (i.e. ignoring iterations before and after) [125]. It is well known that greedy optimisation can be used to build near-optimal subsets of a given set when the objective function is non-negative, monotone, and submodular [49, 125]. Though these conditions are not strictly necessary, results are not so clear when one of these conditions is dropped [49].

While non-negativity is self-explanatory (the objective function cannot be less than zero), monotonicity and submodularity deserve further clarification. In our context, monotonicity will require that cohorts are always more diverse than their smaller sub-cohorts while submodularity will require that an applicant's marginal effect on diversity for a cohort will be (weakly) less than their marginal effect on

[28] A keen reader may note that, under stricter conditions, others have already introduced algorithms for calculating the SPF outright. For example, Kleinberg et al. [82]'s algorithm can be easily extended to calculate the SPF when an organisation only possesses one proportional diversity preference.

diversity for a smaller sub-cohort. More formally, a function D defined on subsets of some universe U is monotone if and only if

$$\forall Y \subseteq U, X \subseteq Y : D(X) \leq D(Y), \tag{7.8}$$

and is submodular if and only if

$$\forall Y \subseteq U, X \subseteq Y, x \in U \setminus Y : D(X \cup \{x\}) - D(X) \leq D(Y \cup \{x\}) - D(Y). \tag{7.9}$$

This may appear constraining, but, luckily, diversity functions $\delta_g^{prop}(c)$ and $\delta_G^{count}(c)$, $D(c)$, the talent function $P(c)$, and $F(D, P)$ all satisfy these conditions as defined in Section 7.3.2. We show this in Theorems 3, 4 and 5 in Appendix C.2.

Now that we have established the necessary restrictions on functions $F(D, P)$, we present a greedy algorithm that finds c to optimise $F(D, P, c, \iota)$; by repeating this for various values of ι, we obtain the frontier between $D(c)$ and $P(c)$ (i.e., the SPF). This algorithm relies on two observations. First, any point on the SPF can be represented as the maximum of a weighted sum $f(\iota, c) = \iota * D(c) + (1 - \iota) * P(c)$ where $\iota \in [0, 1]$. Second, any $f(\iota, c)$ is monotonic and submodular. In this context, the algorithm repeatedly maximises a weighted sum of diversity and talent, varying the weight put on each element in each maximisation. Formally, the algorithm maximises $f(\iota, c)$ m times, where each iteration optimises $\iota = \frac{m_i}{m}$. Then, for each f, this algorithm builds each cohort c from c of size 0 until size n by adding an applicant *not* in the current cohort c ($u \in U \setminus c$) that yields the highest f value (i.e., that maximises $f(c \cup \{u\})$). This algorithm is presented more formally in Algorithm 1.

Algorithm 1 Greedy Frontier Optimisation

For each desired point on the frontier defined by $\iota \in [0, 1]$
  Let $f_\iota := \iota * P + (1 - \iota) * D$ be weighted average of P and D
  Begin with empty cohort $c = \vec{\mathbf{0}}$
  While cohort c is less than the desired size ($|c| < k$)
    Find applicant i such that adding i to c maximises $f_\iota(c + i)$
    $c := c + i$

It is well-known that the greedy algorithm yields a $\left(1 - \frac{1}{e}\right)$-approximation of any submodular, monotonic set function [85]. That is, the algorithm selects cohorts whose f_ι values are at least $\frac{1}{1-\frac{1}{e}}$ of the maximum f_ι any cohort of that size selected from the same applicant pool. For the avoidance of doubt, a proof of these approximation bounds is presented in Theorem 6 in Appendix C.3. Thus, the Greedy Frontier Optimisation algorithm returns points on a curve that $\left(1 - \frac{1}{e}\right)$-approximates the true SPF[29]. We note that this is a worst-case approximation ratio and that the actual approximation ratio may be much better.

7.4 A Field Study with Programme A

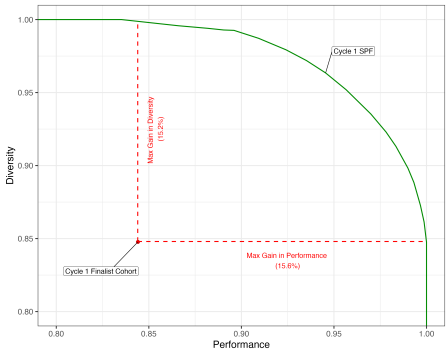
We apply our methodology to evaluate our technology in a field deployment with Programme A's Cycle Y. Through this deployment, we document evidence that Programme A selected finalists within the SPF – consistent with the first prediction of our model – and that Programme A selected much closer to the SPF after they were given an SPF estimate to aid in the selection of their third cohort.

Evaluating Past Selection Decisions Before implementing our technology, we use the methodology described in Section 7.3.1 to determine the efficiency of past selection decisions. In particular, we analyse the finalist selection portion of the Cycles W and X, where the programme must construct a cohort of approximately N applicants from a pool of roughly 4N.

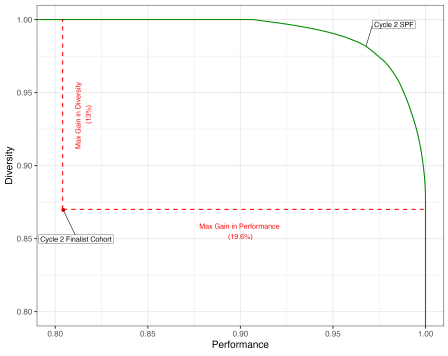
This analysis requires two steps: (1) applying Algorithm 1 to both cohorts to estimate the SPF and (2) comparing the actual talent and diversity levels of the finalist cohort to the estimated SPF. The model we developed in Section 7.3.1 would suggest that this comparison should find that the chosen cohorts are on or near this frontier.

The results from these two steps are depicted in Figure 7.2. Surprisingly, neither the Cycle W nor Cycle X finalist cohorts are chosen on the frontier. (We confirm

[29]In practice, the outputs of the greedy algorithm do not always themselves form a convex curve. We remove produced points that do not sit on the convex curve.



(a) The SPF for the Cycle W finalist selection process. In Cycle W, cohort diversity could have been improved by 15.2% without any reduction in cohort performance, and cohort performance could have been improved by 15.6% without any cost to diversity.



(b) The SPF for the Cycle X finalist selection process. In Cycle X, cohort diversity could have been improved by 13% without any reduction in cohort performance, and cohort performance could have been improved by 19.6% without any cost to diversity.

Figure 7.2: These figures depict the SPFs we estimate for the Cycles W and X finalist selection processes. The y-axis represents the diversity score while the x-axis represents average cohort performance (i.e. project scores). The green curve is our estimate of the cycle SPF, which represents the upper bound of diversity that is achievable at every level of cohort performance. The red dot depicts the actual level of diversity and performance of the finalists that were selected. The vertical and horizontal dashed red lines represent the maximum Pareto gain that was possible along the diversity and performance dimensions respectively. These figures are reproduced at a larger scale in Appendix D.3.

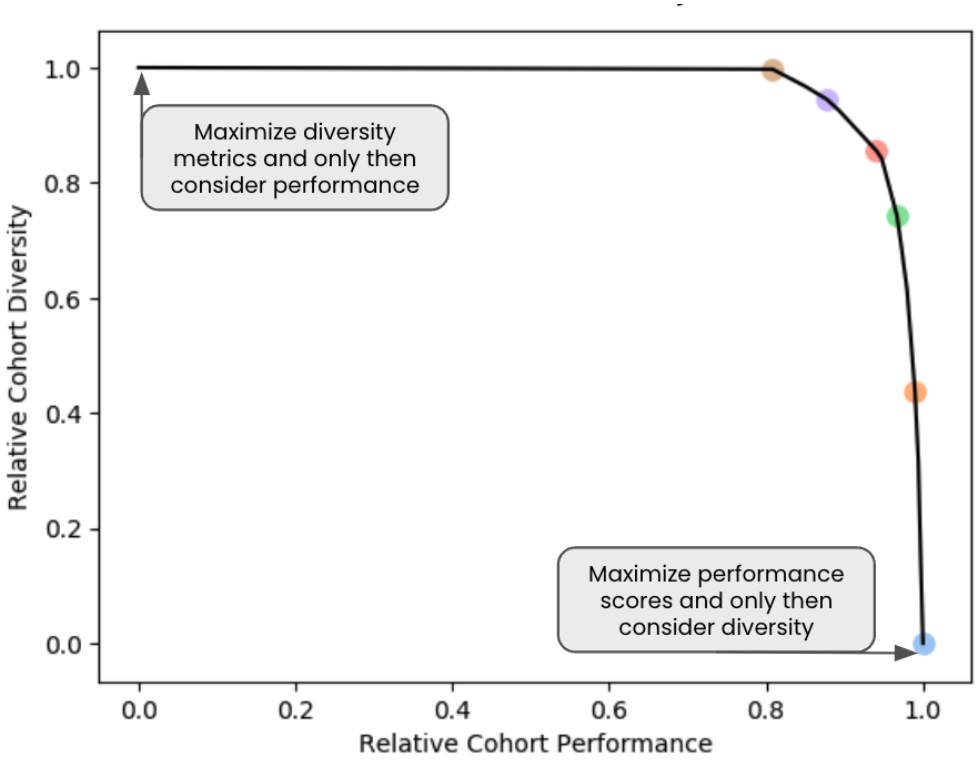
that these apparent gaps between frontiers and chosen cohorts are statistically significant using a permutation test in Figure 7.5.)

These results confound the simple model from Section 7.3.1, which suggests that organisations should always select cohorts on the frontier, as all cohorts within the frontier are dominated by cohorts that are either at least as diverse and more talented or at least as talented and more diverse.

Evaluating Selection Decisions with Decision Support Now we turn to analysing what happened to selection in the talent investment programme when they were given access to our DST in Cycle Y. Again, we first estimate the SPF. However, rather than immediately comparing chosen finalists to this estimate, we instead construct a functional implementation of Prototype 6.3c using this estimate.

Selectors were provided with this DST to inform their decision-making process. The tool, as depicted in Figure 7.3, presented the estimated SPF curve along with

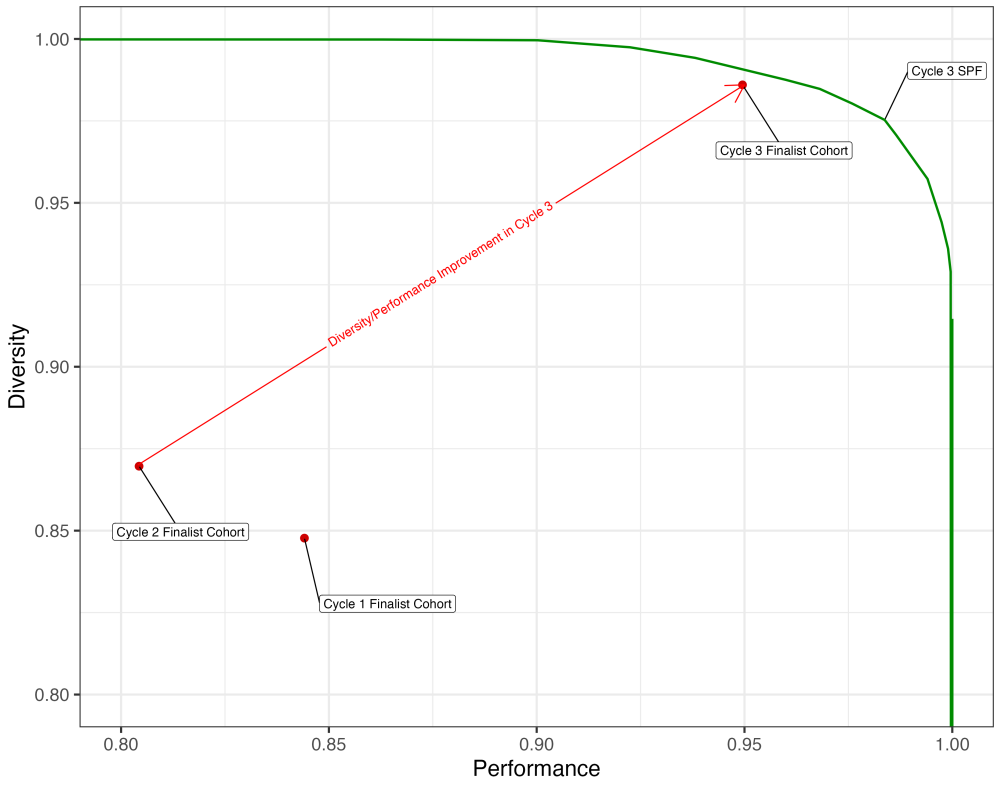
Figure 7.3: This figure displays the SPF-based DST provided to Programme A selectors in Cycle Y. In addition to the SPF itself, selectors were given access to a myriad of supporting information. While this information cannot all be presented here, much of it describes the candidate optima (i.e., the cohorts represented by the coloured dots). In particular, selectors were interested in the spread of performance scores in each cohort, as well as the extent to which each cohort satisfied programme diversity targets.



several pre-calculated candidate cohorts (the "coloured dots"). For each of these candidate cohorts, selectors could review supporting information, including the distribution of performance scores and the extent to which various programme diversity targets were met. This allowed them to evaluate these specific, pre-defined options and understand the associated trade-offs. After selecting a desired cohort along the SPF, this cohort is used to inform a shadow price ι. Participants were then repeatedly shown all available information on the next best applicant to add to the cohort according to chosen shadow price ι and asked to rule that candidate in or out. This process was repeated until the desired size was reached. After the selection process was completed with the aid of this tool, we then compare the actual finalist cohort's diversity and talent to the SPF estimate.

The results from this analysis are depicted in Figure 7.4. Here we see notable differences in the selection patterns relative to Cycles W and X. In particular, the Cycle 3 finalist cohort is nearly on the SPF, making the possible Pareto improvements in both directions no more than 2%. This suggests two things. First, it provides further evidence that selection decisions in Cycle W and Cycle X were, in fact, inefficient; had Programme A known about the possibility of making Pareto improvements relative to their stated preferences, they likely would have changed their behaviour. Second, it provides evidence that the DST presented here actually influences the decisions of selectors.

Figure 7.4: This figure displays the SPF for the Cycle Y finalist cohort. Again, the y-axis represents the diversity score while the x-axis represents average cohort performance, the green curve is our estimate of the SPF, and the red dots depict the actual level of diversity and performance of the finalists that were selected. In this case, we overlay the finalist Cycles W and X cohorts to provide a point of comparison. The diagonal dashed red line represents the distance in diversity-performance space between the Cycle X cohort and the Cycle Y cohort. In Cycle Y, there are no significant Pareto improvements in either diversity or performance.



To determine whether the improvements are statistically significant, we leverage

a permutation test depicted in Figure 7.5. The key comparison is between Cycle Y max Pareto improvements in talent and diversity (the solid blue vertical lines in both panels) and the corresponding 95 percentile of the random difference distributions (the dashed black vertical line). For both dimensions, the possible improvements are statistically insignificant. In contrast, Cycle W and Cycle X both display statistically significant max Pareto improvements. Ultimately, though this confounds the predictions about selector behaviour implied by the model in Section 7.3.1, it does suggest that the DST is effective in improving selection decisions.

7.5 A Plausible Explanation for Selection Inefficiencies

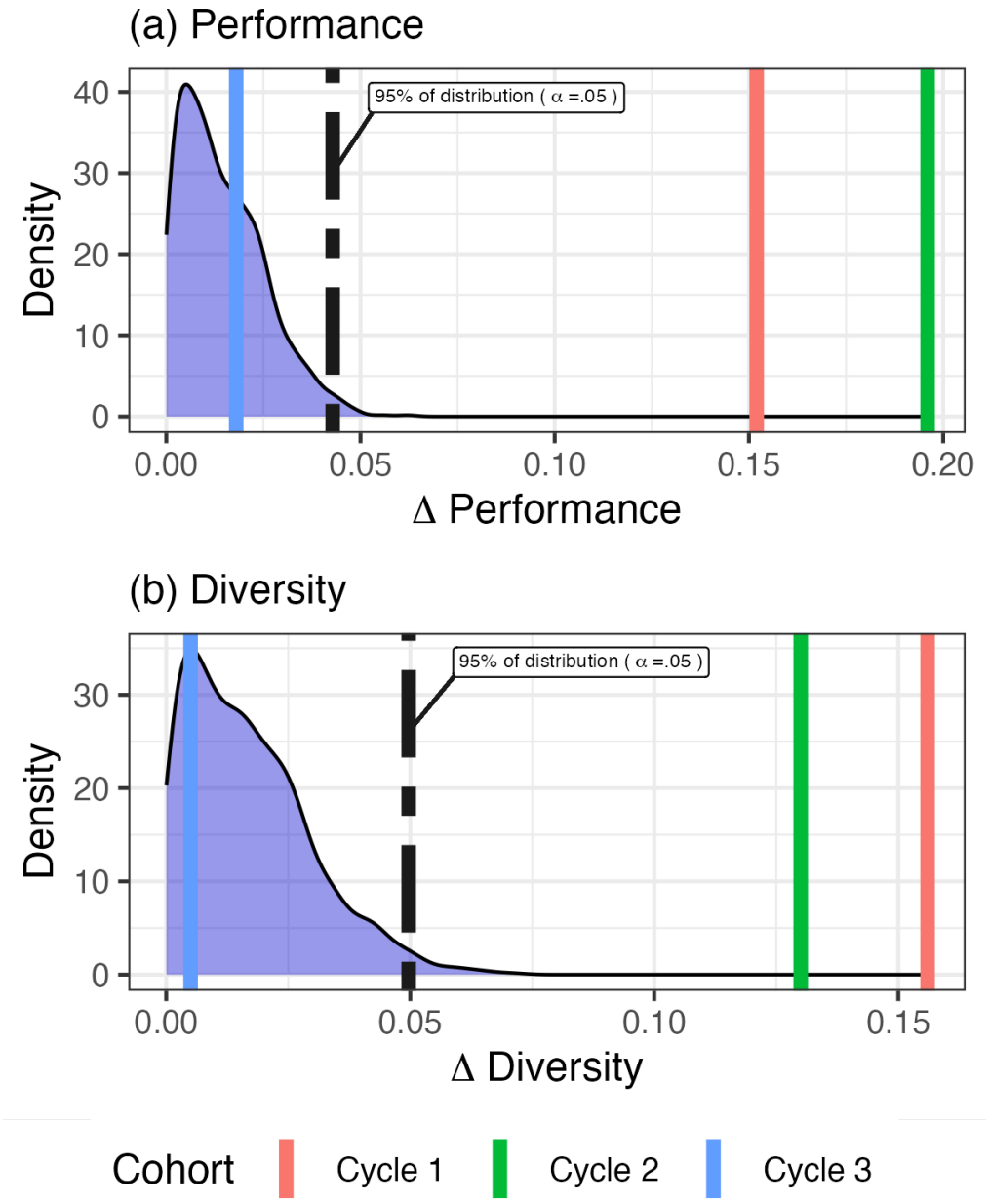
7.5.1 Why Are Organisations Selecting Pareto Inferior Cohorts?

The results of our field study, specifically w.r.t. the Cycles W and X, suggest that organisations are not selecting cohorts on the SPF. This is surprising, as the SPF model suggests that organisations should always select cohorts on the frontier, as all cohorts within the frontier are dominated by cohorts that are either at least as diverse and more talented or at least as talented and more diverse. In conversation with Programme A, we have come to two plausible explanations for why this might be.

First, it might be that the axes of performance and diversity fail to capture the full dimensionality of organisational preferences. One mechanism for this, supported by the revelation from Chapter 6 that programmes desire representations of specific idiosyncrasies, is that our axes fail to capture an idiosyncratic preference possessed by Programme A selectors. Another mechanism, also supported by Chapter 6, may come into play if the organisation's quantifications of talent do not capture the full scope of their concerns; i.e., if part of the measured "talent" an organisation selects for is only captured qualitatively, it cannot be factored into our model.

Second, however, it may be that there is an inherent cost associated with approaching the frontier. This would make organisational selections in the Cycles

Figure 7.5: This figure displays permutation tests comparing the potential Pareto improvements along the diversity and talent dimensions to the distribution of differences on both dimensions from 1000 randomly drawn pairs of cohorts. The dashed black vertical line represents the 95 percentile of these differences. The solid vertical lines represent the maximum Pareto gain on performance and diversity in each application year. We interpret inefficiencies at or larger than the 95 percentile of the distribution as significant, thereby sticking to the conventional α value. While Cycles W and X both appear to have significant inefficiencies, Cycle Y does not.



W and X second-best, in that they are optimal according to a model placing certain costs on their selections.

The selection decisions in Cycle Y favour our second explanation; if the organisation had uncaptured preferences, we would expect these to play a similar role in Cycle Y, and to result in an apparently suboptimal cohort selection. However, the cohort selected is, according to our model, Pareto optimal.[30] Thus, we must seek out the cause of this inherent cost.

7.5.1.1 Diversity Causes Complexity

The most plausible source of cost in approaching the frontier is the impracticality of selection teams discovering the frontier by hand. This is particularly sensible, as we prove here that calculating the frontier outright is **NP**-hard; that is, there are no known efficient algorithms that can calculate the SPF outright [35].

To see why intuitively, consider a college that aims to accept some target fraction of Black applicants from a pool of Black and White applicants. And, assume the school wants to select as talented a class as possible, where talent is proxied for by test scores, grades, or some combination of the two. In this special case, as shown in Kleinberg et al. [82], there exists a computationally easy algorithm to calculate the SPF shown below.[31]

Algorithm 2 A Procedure For Calculating the SPF Based on Kleinberg et al. [82]

Define a minority group (Black) and mutually exclusive majority group (White),
Rank applicants within their group by test score,
Define a target proportion of Black admits,
Select Black applicants from the highest ranking down until the target is reached,
Select the White applicants from the highest ranking down for the remaining slots,
Repeat steps 1-5 for different thresholds of representation to trace out the SPF.

What allows this algorithm to work is the mutual exclusivity of the minority and majority groups. This allows one to transform the diverse talent selection

[30] More specifically, our chosen cohort is not significantly within the frontier.

[31] Technically, the algorithm presented by Kleinberg et al. [82] only optimises for the *most diverse* point on the SPF. Thus, we have added the **Repeat** step (cycling through the algorithm with different representation thresholds) to enable their algorithm to trace out the SPF.

problem into two separate talent maximisation problems where the organisation simply selects the most talented members of each group. This can be extended to any case where target proportions are defined at the level of mutually exclusive group level, even when multiple different demographic dimensions are considered. To be concrete, if the organisation cares about race and gender and, thus, has target proportions for black male, black female, white male, and white female applicants, then the problem can be broken into four separate talent maximisation problems where the most talented members of each group are selected until the target proportions are met for each group.

But what happens if an organisation has preferences for the representation of non-mutually exclusive groups? (I.e., what if an organisation places nonzero weight on two proportional diversity functions?) To continue the running example, this would be analogous to a college that has target proportions for black applicants and female applicants, but not for each race by gender combination. This seemingly small change prevents an organisation from transforming the problem into simpler group-specific talent maximisation sub-problems. To see this, consider applying the Kleinberg et al. [82] algorithm to each group sequentially; this would mean selecting the best black applicants until reaching the target proportion, then doing the same for female applicants. If the most talented black applicants were male or if there were few talented white females in the pool, having allocated the black slots in this way forces the college to select less talented females than optimal (or, it may inhibit reaching the target proportion for females at all).[32] In short, when diversity preferences are over non-mutually exclusive groups, we cannot cleanly and efficiently break the problem into simple talent maximisation subproblems for disjoint minority groups, so it is not clear how we might extend the Kleinberg et al. [82] algorithm to a general version of the diverse talent selection problem.

The general diverse talent selection problem allows organisations to have preferences for the representation of an arbitrary number of overlapping (or disjoint) demographic groups. This aligns more closely with the diversity preferences of

[32]An alternative strategy might iterate over different intersectional targets that satisfy the original two targets; this strategy still suffers from non-polynomial growth.

real-world organisations like colleges, firms, and social impact programmes, many of which aim to select personnel from various ethnicities, genders, classes, geographies, ideologies, and specialities. Organisations generally state their preferences using statements of the following form: "the organisation desires at least $x\%$ of group g" or "the organisation desires at least one person from m groups". These types of diversity preferences are what is formalised in the function $D(c)$, which forms an integral part of the class of functions F; we demonstrate here that calculating F is **NP**-hard.[33]

We now prove that, for the class of functions F of the form from Equation 7.7, the problem of finding the optimal subset of size k for any $f_i \in F$ is still computationally complex. This time, we rely on the assumption that **NP**-hard problems are computationally complex. That is, Theorem 1 holds.

Theorem 1. *Let U be a 'universe' set of size at least $N \geq 2 * n$ and $F = \{f : \mathcal{P}(U) \to \mathbb{R}\}$ be the set of functions described in Equation 7.7. Then $Opt_{spec}(f_i, n) := argmax_{c \in U \wedge |c|=n}(f_i(c))$ is* **NP***-hard in n.*

To do this, and to justify the significance of this result, we bring in the computational complexity of the Vertex Cover problem, which has been proven to be **NP**-hard [35]. Vertex Cover can be seen in Theorem 2.

Theorem 2. *Let $G = (V, E)$ be a graph. Let $VC(G, \kappa) := Cov | Cov \subseteq V \wedge |Cov| = \kappa \wedge \forall e \in E. \exists v \in Cov. v \in e$ be a function of G that returns a set Cov such that every edge in G is incident on at least one vertex in Cov. Then VC in* **NP***-hard in the number of vertices.*

We now prove Theorem 1 by reduction to Theorem 2, assuming that there exists no polynomial time solution to Vertex Cover [35].

Proof. Suppose for a contradiction that Theorem 1 admits some polynomial-time solution Alg_{spec}. I.e., $Alg_{spec}(s_i, k) = argmax_{c \in U \wedge |c|=n}(s_i(c))$.

[33] We do this via 'reduction' to the Vertex Cover. A reduction is simple: $A \leq B$ (i.e., A reduces to B) if and only if there exists a polynomial time algorithm that makes some polynomially bounded number of calls to B and thus returns an answer to A. In other words, we say that A is **NP**-hard if and only if $\forall B \in \mathbf{NP} A \leq B$. It is clear to see, then, that if B is **NP**-hard and $A \leq B$, then A is also **NP**-hard. For more details on reductions, see Papadimitriou [136].

Algorithm 3 An Algorithm for $VC(G = (V, E), \kappa)$

Consider $U := E$
Define $\vec{\mathbf{g}} := \{g_i = v_i \in e | e \in E \wedge i \in |V|\}$ such that each g_i has length E and corresponds to whether an edge is incident on vertex v_i.
Return $Opt_{spec}(1 * D(c, \vec{\mathbf{1}}, \vec{\mathbf{0}}, \vec{\mathbf{0}}, \vec{\mathbf{g}}, \vec{\mathbf{0}}) + 0 * P(c)) \geq k$

Then consider the algorithm Alg_{VC} that is defined in Algorithm 3. But this algorithm solves Vertex Cover in polynomial time relative to Opt_{spec} and thus is a polynomial time solution to Vertex Cover. Assuming $\mathbf{P} \neq \mathbf{NP}$, contradiction! □

7.5.2 Embedding Complexity into the Model

Knowing the **NP**-hardness of calculating the SPF outright, we can more comfortably assume that there exists a search cost in approaching the frontier; knowing that this **NP**-hardness is driven by diversity targets, we can further suppose that this search cost is driven by diversity preferences. This leads us to a new model that incorporates complexity costs into the selection problem.

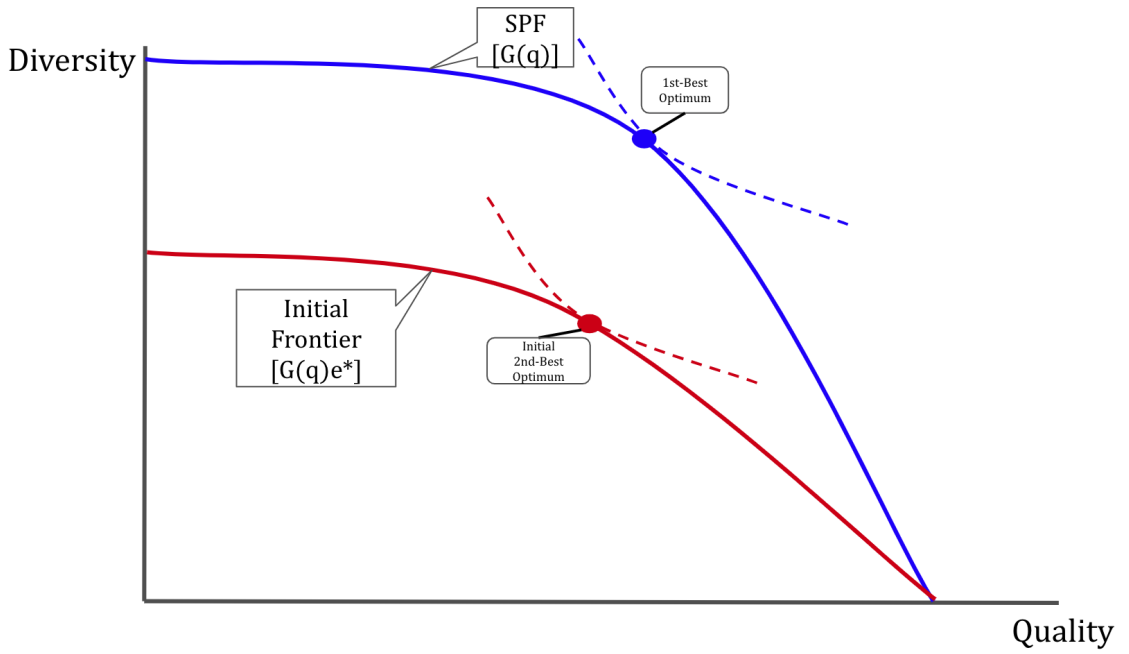
Consider a variation of the simple cohort selection problem (see Section 7.3.1) where the organisation can search for increasingly diverse cohorts at a cost. Let the amount of search effort be $e \in [0, 1]$ and define the cost of search effort to be $\alpha p(e)$ where the cost is convex (i.e. $p_e > 0$ and $p_{ee} > 0$) and α is a constant that is inversely related to the quality of search technology available. Furthermore, we let the amount of search deterministically increase the maximum achievable diversity at each talent level, which is now given by $D^{SPF} * e$. The optimisation problem can then be rewritten as:

$$\begin{aligned} &\max_{d,p,e} F\Big(d, p\Big) - \alpha p(e) \quad \textbf{s.t.} \quad d = G(p) * e, \\ &\implies \max_{p,e} F\Big(\underbrace{G(p) * e}_{\text{Info Cost}}, p\Big) - \underbrace{\alpha p(e)}_{\text{Direct Cost}} . \end{aligned} \tag{7.10}$$

It is clear from the form of the organisation's new objective function in Equation 7.10 that the complexity of maximising diversity imposes two kinds of costs: an information cost that represents the fact that the organisation will generally not

know which cohort is on the SPF and a direct search cost. Also, when search costs are set to zero (i.e. $\alpha = 0$), the problem collapses into the original problem because searching is costless and, therefore, maximised at $e = 1$. But, when $\alpha > 0$, the optimal cohort will now be inside of the SPF. This is because, for all c' such that $D(c') = G(p(c') = p') * e$ there exists c^f on the SPF such that $D(c^f) = G(p(c^f) = p')$. Thus, as long as the optimal effort is below 1, any solution to this problem will result in selecting a cohort that is within the SPF and, therefore, non-first-best. The solution to the selection problem with complexity is depicted in Figure 7.6.

Figure 7.6: This figure depicts an example solution to an iteration of the selection problem with complexity-induced search costs, which is described in Equation 7.10. As in Figure 7.1, the solid blue curve represents the SPF, the dotted blue curve represents the organisation's indifference curve corresponding to the highest achievable utility without search costs, and the blue dot represents the diversity and performance of the first-best solution. Additionally, the solid red curve represents the accessible frontier with optimal search, the dotted red curve represents the highest achievable utility with search costs, and the red dot represents the diversity and performance of the optimal cohort with search costs (i.e. the second-best solution).



Additionally, the extent of the inefficiency will tend to reduce as complexity costs reduce. We can see this by examining the comparative statics of the model. To simplify our derivation of the relevant comparative static, we refer to the organisation's objective function as $O(p, e) \equiv F\Big(G(p) * e, p\Big) - \alpha p(e)$. Furthermore,

we use subscripts on functions to refer to partial derivatives and we drop the arguments of functions where this does not confuse. The (necessary) first-order conditions from this model are, therefore, the following:

$$O_p \equiv F_d(G(p)e, p)G_p(p)e + F_p(G(p)e, p) = 0$$
$$O_e \equiv F_d(G(p)e, p)G(p) - \alpha p_e(e) = 0.$$

To ensure this is a maximum, we also need to assume that the (sufficient) second-order conditions hold. They are the following:

$$O_{pp} \equiv e^2 G_p^2 F_{dd} + 2eG_p F_{dp} + eG_{pp}F_d + F_{pp} < 0,$$
$$O_{ee} \equiv G^2 F_{dd} - \alpha p_{ee} < 0,$$
$$O_{ee}O_{pp} - O_{ep}^2 > 0,$$

where $O_{ep} \equiv O_{pe} \equiv eGG_pF_{dd} + F_dG_p + GF_{pd}$. Under these conditions, solutions to the first-order conditions both exist and guarantee a maximum. These solutions can be defined as $p^*(\alpha)$ and $e^*(\alpha)$. If we plug this into the first-order conditions and take a derivative with respect to α, which governs the complexity costs, we get the following system of equations:

$$O_{pp}\frac{\partial p^*}{\partial \alpha} + O_{pe}\frac{\partial e^*}{\partial \alpha} + O_{p\alpha} \equiv 0,$$
$$O_{ep}\frac{\partial p^*}{\partial \alpha} + O_{ee}\frac{\partial e^*}{\partial \alpha} + O_{e\alpha} \equiv 0$$

where $O_{e\alpha} = -p_e$ and, essential for signing the comparative static, $O_{p\alpha} = 0$. We can then solve for $\frac{\partial e^*}{\partial \alpha}$ algebraically (or using Cramer's rule), which gives the following:

$$\frac{\partial e^*}{\partial \alpha} = \frac{-O_{e\alpha}O_{pp}}{O_{ee}O_{pp} - O_{ep}^2} + \frac{O_{ep}\cancelto{0}{O_{p\alpha}}}{O_{ee}O_{pp} - O_{ep}^2} = \frac{p_e O_{pp}}{O_{ee}O_{pp} - O_{ep}^2} < 0, \tag{7.11}$$

where the the final inequality holds because of the signs assumed in the first and third second-order conditions. Thus, as complexity costs rise, optimal search effort decreases.

This model, thus, implies two predictions about organisational behaviour: (1) when complexity-induced search costs are sufficiently high, organisations will select cohorts within the SPF and (2) as computational costs are reduced, organisations will select cohorts that are closer to the SPF. We have already seen in Section 7.4 that both predictions hold in practice.

7.6 Alternative Applications of the SPF

The main body of this chapter implements and evaluates Prototype 6.3c as an in-process DST by conducting Action Research with the Programme A programme. However, in this section, we discuss potential ex-post applications of the SPF.

Comparing the Diversity Cost of Alternative Talent Measures In some cases, organisations may have multiple alternative talent measures that seem equally valid as measures of individual ability. In this case, the tradeoff between each measure and diversity may help an organisation decide which talent measure they prefer. Two cases where this might be relevant are in hiring and college admissions. In hiring, firms may have multiple measures that predict applicant productivity, but have many ways to weigh the various measures that are roughly equivalent for productivity prediction [64]. This can happen if productivity is multidimensional (e.g., work per hour, tenure, spillovers on others, etc.), and different measures are correlated with some dimensions and not others. A similar problem can be found in college admissions, where, again, the college has multiple measures of applicant ability and may be close to indifferent about some set of ways of combining them when judging an applicant's talent [167].

In these cases, the SPF estimation procedure allows an organisation to consider another dimension: which talent measures demand the sharpest tradeoffs against cohort diversity? I.e.: which measures yield the smallest SPFs? This is particularly

relevant in cases where preferences may be lexicographic, meaning that an organisation wants to maximise talent first, then, conditional on doing so, choose the cohort among top talent cohorts that is the most diverse possible. It also is relevant in contexts where an organisation is not allowed, either legally or internally, to explicitly prioritise diversity in its selection criterion, but still wishes to promote diversity [19].

We demonstrate how to use SPF estimation to compare the diversity tradeoffs of two alternative talent measures. To do this, we estimate the SPF twice, once using each of the talent measures, and then compare the level of diversity at each percentile of both measures. In the case of indifference between the two talent measures on the talent dimension, the measure with higher maximum achievable diversity in the relevant percentile range should be chosen if the organisation cares at all about diversity. We use two measures of talent collected by Programme A: a project-based measure and a traditional score. The results of this comparison are depicted in Figure 7.7. Given that the programme does care about diversity, this would justify using project quality instead of the traditional score for selection.

This method can also be applied to compare the diversity-talent tradeoff across application years. To do this, simply estimate SPFs for each application cycle and compare the level of diversity at each percentile of talent. As long as the diversity goals remain the same each year and cohort diversity is renormalized such that the most diverse cohort across all years becomes 1, organisations can compare across years to see whether differences in applicants across years better afford to get closer to their goals. Figure 7.8 shows just this comparison. In general, the Cycle Y SPF allows for selecting more diverse cohorts at every level of talent than the other two cohorts. But, whether Cycles W or X allow for more diversity depends on where in the talent distribution the programme is interested in. Near the top of the talent distribution, Cycle X has more diverse cohorts, but this flips as talent falls below the 94th percentile.

Evaluating Alternative Selection and Screening Approaches Relatedly, organisations may consider using cheaper, but lower quality measures of talent

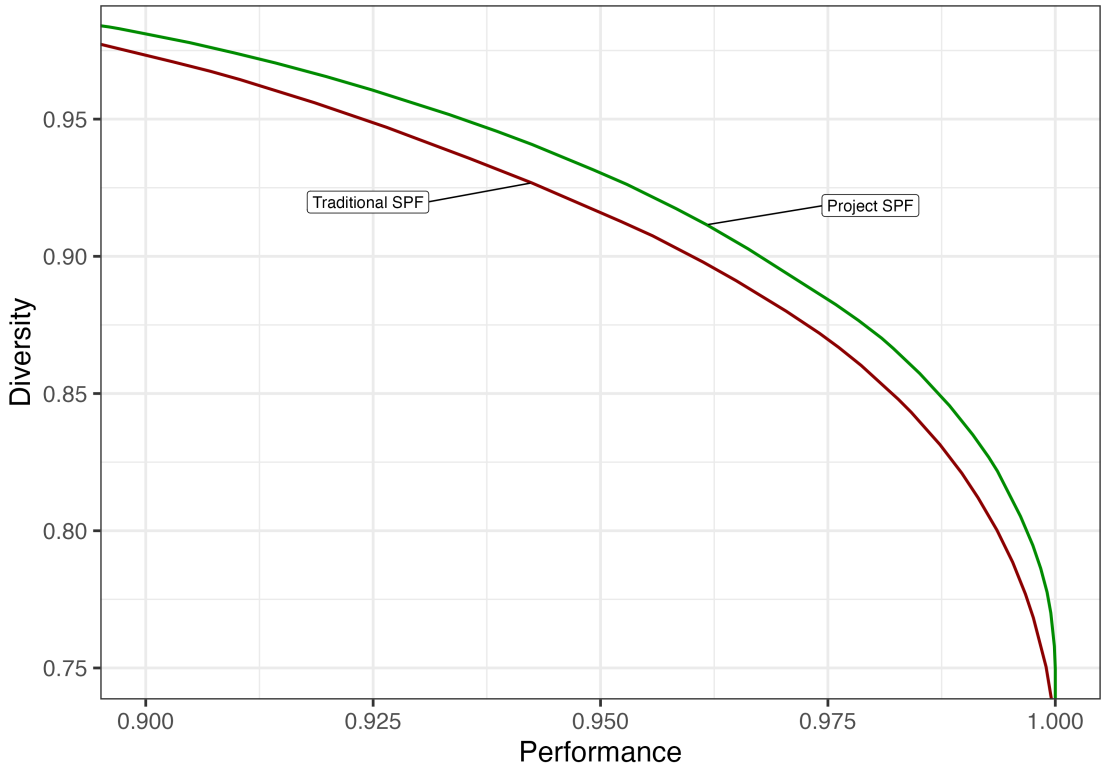


Figure 7.7: This figure displays the SPF we estimated for the Cycle W finalist cohort and an SPF based on a more traditional method of measuring performance (i.e. the average of cognitive ability and an essay assessment). The y-axis represents the diversity score while the x-axis represents the average cohort performance on projects or the traditional score. The vertical distance between the SPFs represents the difference in maximal diversity conditional on a cohort performing at a particular percentile of both scores. We see here that, above the 90th percentile of talent for both measures, the project quality measure strictly dominates the traditional score in diversity.

to screen or select applicants. For example, firms may consider using metrics (e.g., cognitive or personality assessments) or recruiters to screen their applicants rather than allow each applicant to be assessed via an interview. In some cases, organisations may be considering replacing costlier selection measures and selecting applicants entirely based on cheaper information. Unlike before, we now assume that the initial metric captures talent much better than the new metric. Thus, rather than comparing different SPFs, we place cohorts selected using new metrics on the SPF estimate drawn using the original metric. If the new metric is not too much worse than the original metric, then the new metric may be a better choice.

Running selection counterfactuals can be done using two types of designs: the

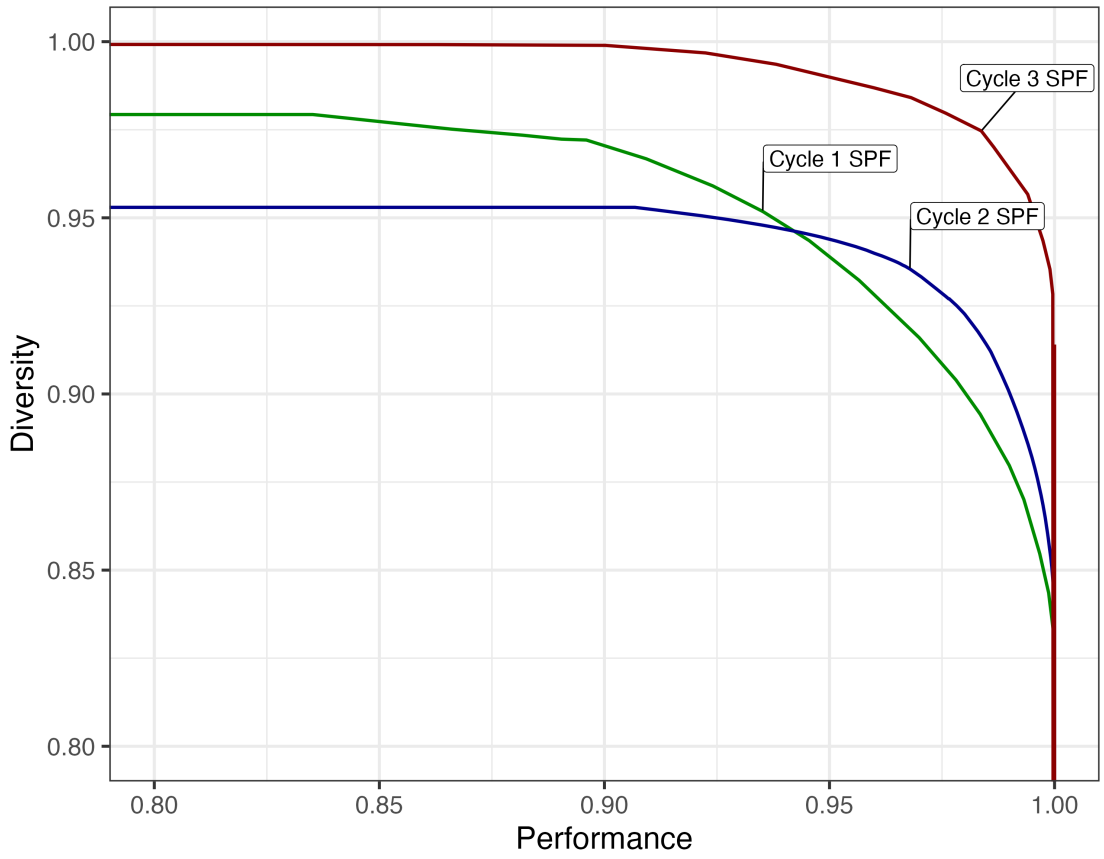


Figure 7.8: This figure displays the SPFs we estimate for three finalist cohorts. The y-axis represents the diversity score while the x-axis represents average cohort performance (i.e. percentiles of mean project scores). The diversity target is held constant across cohorts, so differences in SPFs conditional on performance represent differences in the capacity to reach the same diversity target at a given level of performance.

first we will refer to as a "causal" design and the second we call a "suggestive" design. A causal design requires an organisation to run a screening experiment where applicant talent is either evaluated randomly (or all applicants are evaluated). This allows organisations to avoid the selective labels problem whereby results become biased due to the selection of who gets evaluated and who doesn't. Avoiding this problem allows organisations to analyse representative samples, meaning that comparisons between alternatively selected samples and the estimated SPF should extend to the full population of applicants. Thus, barring any significant contextual changes, the results will be the same (in expectation) when used on another applicant pool (in this sense, the results are "causal"). Alternatively, a counterfactual exercise can be conducted on selected data where only a selected set of individual talents

are assessed. In this case, the applicability of the results to another applicant pool is merely "suggestive", hence the name "suggestive design". Causal designs, though more useful for decision-making, are also more costly, as they require running a selection experiment, which may force organisations to miss out on talent (additionally, using known-inferior selection methods may pose a fairness concern).

To demonstrate, we return to the talent investment programme example where, in Cycle W, we can assess alternative selection procedures using a causal design. This is because, in Cycle W, the programme ran a selection experiment to determine whose projects were reviewed. In particular, the programme used a weighted sum of applicants' cognitive ability and peer assessments of their video essays to select a top tranche who would receive project reviews. Of the remaining applicants, a smaller subset were chosen at random to be evaluated as well. This means that, from the total application pool , a representative sample can be reconstructed by re-weighting the randomly-assessed randomly assessed applicants such that they represent all remaining applicants who were below the project review threshold.

To demonstrate the use of the SPF estimate for counterfactual selection analysis, we compare the efficacy of three alternative selection strategies: the cognitive score, the traditional score, and the peer score. Because the cognitive score and the traditional score both use measures that are closely related to typical talent measures, this comparison also serves as a substantive comparison of traditional selection methodologies and more experimental ones, such as using applicant peer review. The results of this comparison are depicted in Figure 7.9. Here we see that, on the dimension of talent (as measured by project quality) the cognitive ability score performs the worst of the three by far (over 10% worse than the other two scores). The traditional score performs slightly better than the peer score on talent, but the peer score (and the cognitive ability score) performs slightly better than the traditional score on diversity. What is perhaps most striking, however, is that all three alternative selection approaches result in cohorts well within the frontier, indicating that Programme A's actual metric far outstrips each hypothetical alternative.

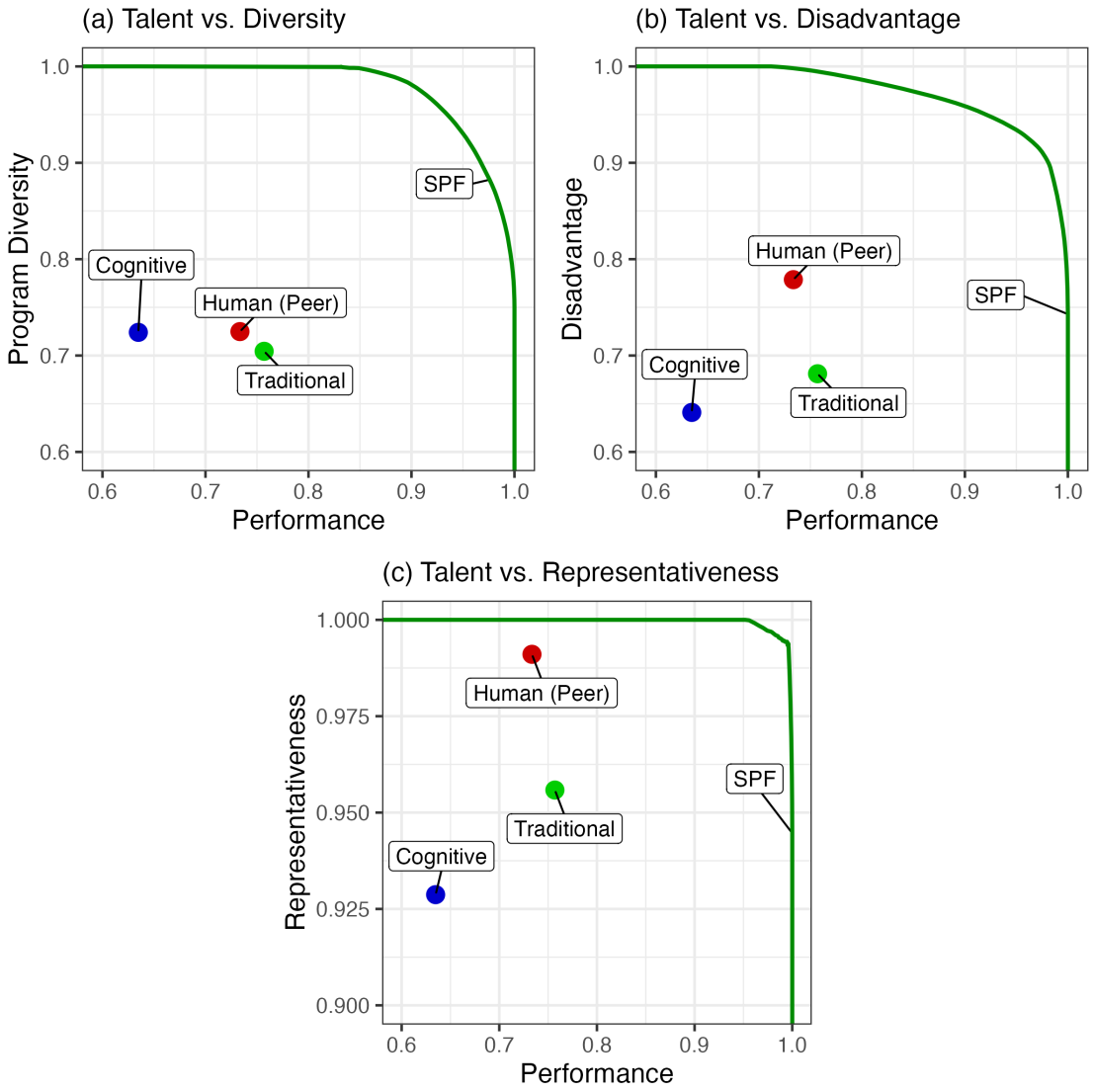


Figure 7.9: This figure displays various SPF estimates for the finalist cohort using the programme's notion of diversity, a 'disadvantage' notion of diversity (drawn from the 'contextualising applications' theme in Chapter 6), and a 'representativeness' notion of diversity (i.e., Prototype 6.3a). The y-axes represent diversity scores while the x-axes represent average cohort performance. The green curves are our estimates of three alternative Cycle W SPFs, which are estimates of the upper bound of diversity that is achievable at every level of cohort performance. Each dot represents the performance and diversity of cohorts had they been selected using only cognitive ability (blue), a combination of written essay judgements and cognitive ability (aka a "traditional" score, which is green), and just peer review (red).

Evaluating According to Alternative Notions of Diversity A similar process can help organisations understand the practical implications of different kinds of preferences over types of diversity. While we isolate three themes relating to

definitions of diversity in Chapter 6, we note that the Programme A programme has a working understanding of what they mean by diversity, and did not wish to adopt any of these notions. In practice, Programme A's diversity targets suggest both 'representativeness' and 'contextualising applications' (which they internally call 'disadvantage' or 'boostability', variously) as important to their consideration of diversity, while 'different perspectives' do not appear in their decision-making process. Thus, Figure 7.9 also depicts results using two alternative notions of diversity based only on the disadvantage and representativeness portions of the Programme A targets. The disadvantage diversity score puts maximal weight on representing those from various historically disadvantaged groups (e.g., being first-generation, poor, or female) while the representativeness diversity score only uses proportional targets that match the demographic distribution of the applicants. Evaluating each alternative selection method indicates that, while selecting on peer judgements or the traditional score both do substantially better than using cognitive ability alone on the talent dimension, using the peer score is by far the highest performing on both disadvantage and representativeness.

7.7 Conclusion

While Chapter 6 focused on the theoretical and empirical aspects of diversity, this chapter has focused on the practical implications of diversity in selection. In doing so, we have introduced the notion of a selection possibilities frontier via a simple model of diverse talent selection, have implemented Prototype 6.3c as an in-process DST, and have demonstrated its use in practice. Analysing decision-making with and without our DST, we have shown that Programme A selects talented and more diverse cohorts when given access to our SPF-based DST. To explain why, we showed that the diverse talent selection problem is **NP**-hard and augmented our model with a notion of complexity costs; this new model predicts that organisations who are better and more cheaply able to approximate the frontier should find themselves closer to it.

Finally, we have also shown that the SPF can be used ex post to compare the diversity tradeoffs of alternative talent measures, evaluate alternative selection and screening approaches, and evaluate according to alternative notions of diversity.

In an age of rapidly expanding interest in selecting from diverse talent pools (signalled by the growth of DEI), this chapter has wide-ranging policy implications. First, the chapter suggests that organisations will face extreme difficulty achieving their diversity goals unless they are willing to adopt more sophisticated selection technology. Second, this chapter contributes methods that are particularly useful for assessing the diversity impacts of alternative merit-based selection strategies. This extends beyond selection to related contexts like hiring, where not appropriately considering the diversity implications of selection strategies can result in lawsuits, and U.S. university admissions' non-merit-based selection has become a legal grey area despite university commitments to diversity.

8

Discussion

Contents

8.1 The Role of AI Systems in Selection: From Decision Support to Selection-Oriented AI

8.1.1 Current Challenges in AI for Selection

8.1.1.1 To Support or Supplant?

AI tools have long been posed as both replacements for and supports to human decision-makers, both within and outside selection processes [12, 66, 73, 92, 184]. Naïve implementations of AI often supplant human decision-makers, sometimes to disastrous effect [92]. As such, we must be careful to design AI systems that support, rather than supplant, human decision-makers.

8.1.1.2 Who to Support?

More human-centric AI systems, such as Explainable AI (XAI), often serve as Decision Support Tools (DSTs) that place the stakeholder—in this case, the selector—at the core of the decision-making process. These systems focus on enhancing the experience of human decision-makers, often by satisfying their subjective desiderata. However, as Lipton [100] argues, post-hoc explanations designed to satisfy these desires may prove more misleading than insightful. In Chapter 4, we extend this critique to post-hoc justifications more broadly and argue it applies to all AI DSTs that prioritise user satisfaction over objective outcomes.

Furthermore, selection involves a complex interplay of interests. Organisations aim to identify the best possible cohort, while candidates seek to be selected. These two interests are often at odds. Beyond these immediate stakeholders, society has a vested interest in ensuring that selection processes uphold values like fairness, diversity, and justice. In pro-social programmes like global scholarships, society also has an interest in ensuring that the most deserving and impactful scholars are chosen.

Centring DSTs solely on the needs of either applicants or selectors would not only suffer from the problem of subjective desires identified by Lipton [100] but would also fail to address the broader social implications of selection. Thus, there

is a need for DSTs that orient themselves around the social values of selection itself, rather than the preferences of any single stakeholder group.[34]

8.1.1.3 What to Support?

The conventional DST paradigm often assumes a series of similar decisions on different cases, such as deciding whether to grant a loan [13, 37, 144, 148, 171]. However, the decision points we enumerated in Table 2.1 are not all captured by this model. The conventional paradigm focuses on a specific type of in-process decision, excluding other in-process decisions and all ex-post decisions. A paradigm for selection-oriented DSTs should therefore seek to support a wider range of decision types.

8.1.2 Proposing a New Paradigm: Selection-Oriented AI (SOAI)

In response to these challenges, this thesis proposes a novel paradigm: Selection-Oriented AI (SOAI). SOAI reimagines the role of AI in talent identification, advocating for a shift away from a purely human-centric framework toward a hybrid selector-centred and selection-driven approach. In this paradigm, the design of AI systems is grounded in the social values that selection processes ought to uphold, with those values supplied and interpreted by the selectors making the decisions. While selectors remain the primary users, they are not the sole focus. Instead, SOAI emphasises evaluating the broader social goals of selection and helping selectors achieve them [53].

The shift toward SOAI represents a necessary evolution in AI for talent identification. By prioritising the social values that selection processes ought to reflect, SOAI challenges the current practitioner-centred approach and introduces a new standard for evaluating the success of AI in decision support. It offers a path toward

[34]We recognise there that selectors' preferences often overlap with these social goals, while other stakeholders may have conflicting preferences. Furthermore, as selectors form the relevant decision-making processes, achieving the social values of selection often requires satisfying the preferences of selectors. Thus, while we centre the DSTs around the social values of selection, we also frequently consider the preferences of selectors.

creating AI systems that not only assist in identifying talent but also ensure that the selection process itself is fair, just, and inclusive.

8.2 Design Recommendations for SOAI Designers

8.2.1 Design for Specific Social Values

While we wish to design to support all social values in the selection process, Chapter 6 demonstrates the difficulty of unpacking the social value of diversity; in Chapter 6, we find success instead focusing on smaller component values that comprise diversity. Similarly, each decision point supported in Chapter 5 implies a specific ontology about the role of generative AI in selection; these, too, stem from specific social values promoted by an organisation.

There is support for this from the literature. For example, literature on algorithmic fairness has long wrestled with contradictions between measurements of different kinds of fairnesses [79]. While 'individual fairness' draws on procedural notions of justice to ensure that applicants are treated equally regardless of differences in protected or irrelevant characteristics [46], 'group fairness' draws on distributive justice in seeking to achieve parity between different demographic groups [33, 129]. There exists literature attempting to reconcile these notions: Zemel et al. [186] attempt to reconcile this in practice by simultaneously optimising for multiple fairness metrics; Lahoti et al. [89] seek only to optimise for individual fairness, and yet find increases in group fairness; and Binns [17] contends that standard, blunt implementations of individual fairness should be replaced with a more nuanced formulation compatible with group fairness. Nonetheless, as they are often implemented, these two notions of fairness are often in conflict, and though designing to support both may be possible, it is liable, in a scholarship context, to create unclarity of the sort plaguing diversity, impeding programme desire to assess these concepts with specific targets.

We suggest this generalises to SOAI practices in general: rather than designing around myriad values, only to find conflicting design implications of these disparate

values, designers seeking to support social values in selection processes should focus on specific social values worthy of consideration.

8.2.2 Identify Decision Points with the Decision Matrix

In Chapter 2, we conceive of selection as a series of decisions. Chapter 5 expands on this, introducing the Decision Matrix framework to categorise the many decision points that selectors face according to their two most germane axes: the stakes of the decision and its stage in selection. This framework allows designers to reason about groups of decision points in much the same way that the explainable AI community reasons about groups of explainability techniques and to isolate desired or required properties of DSTs based on the taxonomy of the decision point they seek to support and to then determine which categories of decisions different GenAI detectors are suitable to support[44, 51, 55, 88, 117].

In Chapter 4, we respond to criticisms isolated to Friedrich and Zanker [55]'s 'post-hoc' explanations; here, the taxonomic distinctions are used in criticism to expand the scope of individual critiques [13, 88]. We suggest the Decision Matrix can be used similarly, to discuss and critique decisions in a scholarship context.

However, we caution designers following this design recommendation to ensure that they also follow design recommendation 8.2.4 and evaluate real change in addition to subjective desiderata. The Decision Matrix framework can be used to derive a set of a necessary, but perhaps not sufficient, properties of DSTs.

8.2.3 Balance Qualitative and Quantitative Information in Presentation

Human decision-makers often desire both a qualitative understanding of applicants and quantitative metrics to compare them. In Chapters 4 and 6, we find that selectors from Programme A and Programme B seek to make decisions informed by both kinds of information; despite this, the desired balance between these modes of information varies based both on practitioner and type of decision. When quantitative information is neglected, practitioners are forced to make decisions on

a case-by-case basis without important numerical context comparing applicants to a larger group; when qualitative information is neglected, practitioners are unable to consider applicants holistically. Developers following SOAI should consider the balance between quantitative and qualitative information in their systems, and design their systems to provide both when necessary.

We again find parallels in the fairness literature to this balance. Qualitative information enables the selectors' consideration of applicants as individuals, and this combines with the process of holistic review to create full pictures of applicants. [46]'s individual fairness holds a similar lens; rather than looking at applicants in terms of their place in the cohort, this notion of fairness demands equal treatment of applicants as people. However, quantitative information makes possible considerations of distributive notions of justice and group fairness principles [129], as decision-makers have access to the supporting information needed to contextualise applications relative to other members of protected groups. Notably, programmes with different ontologies governing what they consider fair will thus have different preferences considering the balance of quantitative and qualitative information in their systems. (This relationship is not absolute, though, as other differences in programmes may lead to differing priorities.)

8.2.4 Evaluate Real Change in Addition to Subjective Satisfaction

Lipton [100] critiques explainable AI (XAI) systems because they risk satisfying the subjective desires of the users while failing to improve objective outcomes. In Chapter 4, we confirm that this critique applies to some post-hoc justifications of model recommendations, as the justifications were found to yield an unwarranted increase in trust in human decision-makers. Thus, it is important to define and evaluate measures of the social values that DSTs intend to support; when evaluating these DSTs, they should not be evaluated human-centrically (i.e., according to their users' satisfaction), but should instead be evaluated on whether their employment improves social outcomes.

The fundamental challenge with evaluating "real change" in a selection context is the lack of ground truth; i.e., there are not, in general, "correct" or "incorrect" selection decisions, only those preferred by the organisation. In this thesis, we solve this problem by working with programmes to define measurable criteria that act as a surrogate for "correct selection decisions". In Chapter 4, these criteria are arrived at through an Action Research (AR) process and expressed in Figure 5.3, while in Chapter 7, these criteria are supplied directly by Programme A, as the programme has internal metrics for both axes of the SPF. We recommend designers work with organisations to define surrogate criteria that can be used to evaluate the success of their systems.

8.3 Implications

8.3.1 Algorithmic Fairness in a Selection Context

The work in this thesis has implications for the broader discussion of algorithmic fairness in selection processes. The design of SOAI DSTs has the potential to impact the lives of many of the world's most vulnerable people; it is thus imperative that these systems are designed fairly. However, the notion of fairness itself is complex and multifaceted. As Kearns et al. [79] highlight, fairness can be understood in both procedural and distributive terms, and different methods of achieving fairness across different subgroups often conflict. Individual fairness is often discussed in the algorithmic fairness literature [46]; this is often contrasted with "group" fairness [10, 17, 50, 52]. Despite attempts to reconcile these differing notions of fairness, such as those by Binns [17], contradictions remain between metrics used to measure different forms of fairness; that is, decisions that may be ruled more fair by certain individual or procedural fairness measures might create group or distributive unfairness. What's worse, scholars disagree even on the best implementations of notions of fairness [17, 52], and differing interpretations conflict.

It is worth noting, then, that the work on generative AI detection in Chapter 5 is built on a desire for procedural fairness, while the diversity goals of Chapters 6 and 7, in practice, accord closely with distributive notions. This raises the possibility

that, via SOAI methods, researchers could determine socially beneficial fairness metrics to uphold in DSTs and build to support those.[35]

8.3.2 New Developments in AI for Selection

The growing popularity of GenAI has already dramatically increased the number of applications that job, university, and scholarship programmes must select from [76]. While a blanket ban on GenAI in application-writing may solve this [170], we find in Chapter 5 that such a ban is unenforcible at present. We note in Chapter 5 that our research is complicated by the rapidly changing nature of both GenAI and detectors. Here, we extend this complication to SOAI as a whole. It may be that, as GenAI development moves beyond retrieval-augmented generation to more complex architectures [97], such as integrated reasoning systems or agentic AI [161], these systems will once again fundamentally change the process of selection. In light of this, SOAI is necessary to ensure that new, more powerful AI systems further the social aims of selection processes.

Of particular interest would be the development of AI systems capable of encoding domain knowledge in their structures, which could support decisions in a fundamentally different way. This could be particularly useful in automated essay scoring, where domain-specific knowledge is a significant problem [106]. If this is the case, then the work done in this thesis may serve as a precursor to the development of these systems and a guide for how to ensure that these systems are designed to support the social aims of selection processes.

8.4 A Critical Reflection on the Position of this Research Within Structures of Power

8.4.1 Critical Assessment of Research Contributions

While this thesis proposes SOAI as a paradigm for more socially conscious AI design, it is essential to critically examine its potential unintended consequences,

[35]This work, in particular, should not be done solely from the decision-maker's perspective. Marcinkowski et al. [105] investigate applicant perceptions of appropriate fairness metrics; this work may be a good starting point for SOAI work in this field.

particularly those arising from the approach to diversity measurement and optimisation in Chapter 7.

8.4.1.1 The Risk of Quantitative Overemphasis and Tokenism

The Selection Possibilities Frontier (SPF) framework developed in Chapter 7, while intended to balance quantitative and qualitative information, risks amplifying the tendency to quantify human worth. By creating numerical representations of diversity and performance, we risk reducing complex human experiences to algorithmic inputs. This is particularly concerning because, as we argue throughout this thesis, qualitative components are crucial for understanding an applicant's lived experiences, contextual challenges, and unique perspectives. They provide the scaffolding that allows selectors to understand not just what an applicant has achieved, but how and why those achievements occurred. Our framework, if misused, could encourage selectors to treat these vital qualitative insights as merely supplementary.

Perhaps more critically, the diversity measurement approaches developed here risk enabling a sophisticated form of tokenism. By providing tools that allow programmes to demonstrate measurable improvements in diversity metrics, these systems may satisfy an organisational desire to appear inclusive while failing to address the deeper structural inequalities that cause exclusion. The SOAI paradigm, while oriented toward social values, operates within existing institutional frameworks rather than challenging them. As Ahmed [1] argues, "diversity work" can create the appearance of inclusion while leaving fundamental power structures intact. Our tools could inadvertently become part of this dynamic, helping powerful institutions deflect calls for more fundamental reform by pointing to their use of socially conscious AI as evidence of their commitment to equity.

8.4.1.2 The Convergence Problem and Algorithmic Monoculture

A second major concern is the risk of algorithmic convergence. If multiple scholarship programmes adopt similar SPF-based approaches, they may begin selecting for overlapping pools of candidates who excel according to the same quantitative

metrics. This could create a new form of algorithmic bias where certain types of applicants—those who perform well on the specific measures captured by these systems—receive disproportionate opportunities, while other forms of excellence are systematically undervalued. This algorithmic monoculture would undermine the very diversity goals these systems are designed to support.

This risk is amplified by the tendency of machine learning systems to optimise for what is easily measurable. If all programmes converge on similar optimisation targets, we risk a narrowing of what is considered valuable, systematically disadvantaging those whose strengths lie outside these predefined frameworks.

8.4.2 Broader Implications for Power Structures

In a seminal piece, Barocas et al. [11] ask whether algorithms challenge or reinforce existing power structures. In the case of this thesis, the answer is complex. While SOAI seeks to improve fairness, it operates fundamentally within existing institutional frameworks that themselves can perpetuate inequality. The scholarships examined here—funded by major philanthropic organisations—provide invaluable opportunities for individuals but also serve to entrench their funders in institutional power structures [188]. By designing AI for these programmes, we may be contributing to the legitimation of these structures.

The generalisability of SOAI to other high-stakes contexts like hiring and university admissions raises additional concerns. Widespread adoption could amplify the risks identified above, potentially creating a 'selection-industrial complex' where a narrow set of algorithmic approaches dominates opportunity allocation across society.

8.4.3 What Could Go Wrong: Key Failure Modes

Several specific failure modes could emerge from the widespread adoption of the approaches developed in this thesis:

- **Metric Gaming:** As programmes become more transparent about their metrics, applicants may develop strategies to game them, undermining their

validity and creating new advantages for those with the resources to understand and exploit the system.

- **False Precision:** The mathematical sophistication of the SPF may create an illusion of objectivity that masks the subjective value judgments embedded within it, making it harder to critique or adjust these systems when they produce problematic outcomes.

- **Institutional Complacency:** By providing tools that demonstrate measurable progress on diversity metrics, these systems may reduce the pressure for more fundamental, and more difficult, institutional reforms.

8.4.4 Guarding Against Failure: Deliberation, Contestability, and Iteration

Given these critical risks, SOAI systems must be designed not as static, authoritative solutions, but as evolving tools that support ongoing human deliberation and are open to contestation. This requires several safeguards:

- **Transparency in Values and Trade-offs:** Systems must make their embedded values (e.g., specific definitions of diversity) and trade-offs transparent to selectors. This includes clearly articulating how different metrics are weighted and how qualitative information is incorporated or potentially sidelined.

- **Mechanisms for Contestation:** Selectors and other stakeholders must have avenues to question, critique, and challenge system outputs. This could involve features that allow users to flag problematic recommendations, suggest alternative interpretations of data, or adjust system parameters under controlled conditions.

- **Support for Iterative Refinement:** SOAI tools should be built with the expectation that they will require regular review. This includes designing for the easy updating of models and metrics as organisational goals evolve or as unintended consequences are identified.

- **Interfaces for Qualitative Nuance:** To counteract quantitative overemphasis, interfaces must actively encourage the integration of qualitative nuance. This might involve dashboards that juxtapose quantitative scores with rich qualitative summaries or tools that help selectors document and weigh contextual factors that are not easily quantified.

Designing for deliberation, contestability, and iteration aims to foster a more responsible and adaptive use of AI in selection. It positions SOAI not as a replacement for human judgment but as a catalyst for more informed, reflective, and ethically aware decision-making. This approach acknowledges that achieving real change requires continuous engagement with the complexities and potential pitfalls of AI-driven selection.

8.4.5 Reconciling Critique with Pragmatic Impact

Despite these significant concerns, this research operates within a pragmatic context. Without it, the institutional structures examined would continue their existing selection processes, likely with less information and greater inconsistency. While this work does not dismantle the structures that concentrate opportunity, it does seek to improve the fairness and efficacy of decision-making within them.

The tension here is between the perfect and the better. While an ideal solution might be to entirely restructure how opportunity is allocated, the practical reality is that these institutions persist. Given their persistence, there is value in ensuring the opportunities they provide are distributed as fairly and effectively as possible, even while acknowledging that such improvements may inadvertently legitimise the broader system. Future work must actively monitor for the failure modes identified here and be accompanied by transparency about the limitations of these systems and the structural inequalities they cannot resolve.

8.5 Limitations

In scenarios outside selection, HCI research often seeks to harmonise the needs of different user groups. In talent identification, however, the conflict of incentives between applicants and selectors is inherent. Our solution has been to centre the organisations, using their stated preferences and past data to evaluate their decisions. This positions the research from the perspective of the decision-makers and may limit the social benefit by not fully engaging with the perspectives of decision subjects.

We similarly assume that the broader social aims of selection align more closely with practitioners than with applicants. However, there is evidence that applicants themselves value fairness and may even prefer algorithmic decision-making in some contexts [105]. By positioning SOAI as a paradigm for DSTs rather than for fully automated systems, we may be limiting its potential impact if algorithmic decision-making proves to be more effective at achieving social goals.

Finally, while the Decision Matrix framework is a useful tool, it intentionally elides certain distinctions to focus on stage and stakes. This simplification, while developed in concert with our partners, may not generalise to all contexts. Furthermore, our focus on improving selection within existing applicant pools means we do not address the broader issues of access and outreach that determine who applies in the first place—a significant limitation, given that the most profound inequities often occur before the formal selection process begins.

8.6 Future Work

While the work in this thesis articulates SOAI as a paradigm for all AI design oriented around supporting selection problems, we develop and test this paradigm for three specific families of decision points. A straightforward extension of this work would apply SOAI principles to other decision points in selection processes. Natural candidates include: supporting essay judgements with automated essay scoring, where a large body of literature already seeks to score applicant essays via algorithm [36, 106, 147, 178], but automated approaches continue to struggle with

marking top or bottom essays [106]; supporting testing and test evaluation with automated scoring systems [34, 131]; and supporting pre-application portions of the outreach process, which was requested by several participants in Chapter 6.

Though SOAI, as we investigate it here, aligns most closely with the interests of selectors, there is a need for human-centric work seeking to determine applicant perceptions of positive social outcomes. While work exists examining applicant perceptions of decisions made about them, [69, 135], this work often approaches research from a fairness or decision-subject-empowerment perspective. No work exists approaching applicant perspectives from the perspective of the ultimate social benefit of selection. Future work should seek to understand how applicants perceive the social outcomes of selection decisions, and how these perceptions can be used to design more effective AI systems for selection.

Though the Decision Matrix provides a useful framework for categorising decision points in selection processes, it is not exhaustive. Future work should seek to augment the Decision Matrix with additional axes that capture the complexity of selection decisions more fully. In particular, though selectors drew a distinction between individual- and group-level in-process decisions in Chapter 6 (and though the design prototypes reflect this distinction), the Decision Matrix does not capture this distinction. Future work should seek to augment the Decision Matrix to distinguish individual- from group-level distinctions and implement more individual-level decision support systems in practice.[36]

8.7 Conclusion

In this thesis, we pioneer a new paradigm of AI design for selection processes, Selection-Oriented AI (SOAI). Chapters 4 and 5 find that existing AI systems often fail to meet the needs of selectors, particularly for in-process decision-making. In response, we propose a new paradigm, SOAI, which seeks to centre the design of AI DSTs not around the selectors but around the social aims of selection they seek

[36] Chapters 5 and 7 both avoid individual-level implementations in real decision-making pipelines due to risks associated with introduced unfairness or bias [10, 14, 64, 79, 99]. Any work implementing tools at the individual-level should consider these risks first.

to practice. Chapters 6 and 7 apply SOAI principles to design DSTs to support considerations of diversity in selection; we implement a prototype designed to satisfy selector desires and find that it improves both diversity and performance outcomes in selection. We then provide a set of design recommendations for SOAI designers, including focusing on specific social values, identifying decision points with the Decision Matrix, balancing quantitative and qualitative information, and evaluating real change in addition to subjective satisfaction.

More broadly, the use of SOAI to support scholarship-specific selection decisions implies the potential to support and improve related decision-making processes, from other selection contexts (e.g., admissions or hiring) to non-selection decision-making contexts (e.g., programme outreach). With a technology-induced flattening of the world [54], more candidates from more parts of the world find themselves qualified for opportunities. Add to this the ease of application submission created by GenAI assistants, and it is clear why applications to job, university, and scholarship opportunities have seen a dramatic increase in recent years [76]. In light of this, we conclude with a call for SOAI across selection contexts; the need has never been more pressing.

Appendices

A

The Programmes we Study

Contents

A.1 Foreword to Appendix A

We work with two global scholarship and talent investment programmes (Programme A and Programme B). Both programmes have asked that they not be identified in public-facing research, and thus this version of the thesis redacts identifying details (programme names have been replaced with pseudonyms; funders, URLs, exact cohort sizes, dates, and programme-specific assessment tool names have been removed or generalised). Methodologically relevant content is preserved.

Table A.1: This table enumerates relevant measurement categories from Programme A and Programme B.

Measure Type	Programme A	Programme B
Cognitive Assessment	ICAR-based test; gamified skills test	ICAR-based test; divergent-thinking task
Essay Review	Peer; External Expert	AI-assisted; External Expert
Grade and Achievement Review	None	External Expert
Finalist Activity Review	Selector	Unknown

A.2 Programme A

A.2.1 Programme Overview

Programme A is a global scholarship and talent investment programme that finds and selects talented and disadvantaged young people from around the world and helps them achieve their full career and service potential. Programme A supports selected scholars and finalists with a variety of benefits accessible at different points in their lives.[37] We work with Programme A across several application cycles, during which time the programme has selected several hundred scholars and several thousand finalists from hundreds of thousands of applications.

Programme A uses a flexible benefits model, where scholars (and, in some cases, finalists) gain access to a variety of potential resources, but utilise only resources they demonstrate a need for. Programme benefits include academic scholarships, educational resources and programmes, networking opportunities, and funding for scholar-led startups.

A.2.2 The Selection Process

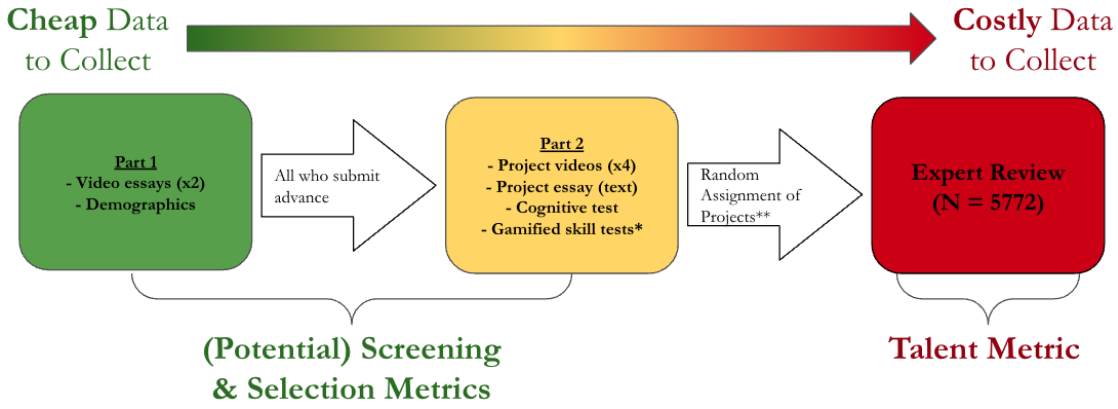
The programme uses a two-stage selection process designed to be accessible to candidates from various global and socioeconomic backgrounds. In stage one, applicants submit various application materials asynchronously; Programme A selects finalists based on the quality of those materials and the programme's cohort

[37]We adopt the terms "scholarship and talent investment programme" and "scholars" as generic, programme-anonymising language; the programme uses programme-specific terminology that has been redacted here.

composition goals. In stage two, finalists engage in one of several synchronous assessment events consisting of various collaborative live activities and an interview; after all events are completed, Programme A uses information from both stages to select scholars.

Stage one of selection occurs asynchronously and in two parts. The first part requires applicants to submit an application form with their demographic information and a small number of essays (in video or written form) speaking to their motivations and background. In the second part, applicants complete a set of digital cognitive assessments and a multi-stage project component documenting work the applicant has undertaken. Specific prompts and the structure of the project component have been generalised to preserve programme anonymity. For an overview of the stage one selection design, see Figure A.1.

Figure A.1: This figure schematizes the key elements of the talent investment programme's data collection and selection process.



Stage two of selection occurs synchronously (though still remotely) in one of several assessment events. Each event consists of up to five activities of three types: presentations (where finalists present information about their project), group activities (where finalists collaborate to discuss and solve problems), or interviews (where finalists are interviewed). All activities were judged by a pool of adult 'selectors' who assessed finalists according to a rubric. Scholar selection decisions were made based both on data collected in stage two and information retained from stage one.

A.2.3 Data Collection

Across the application cycle, Programme A collects a variety of data from applicants. This data includes traditional merit-based measures – including cognitive tests, written essays, and referrals – as well as non-traditional measures – including peer-reviewed video essays, gameified skill tests, and application platform behaviours. We discuss methodologically relevant measures here. More detail on these measurements can be found in Chapter 7, where findings depend on the specifics of Programme A programme measurement. A comparison of the measurement categories used by Programme A and Programme B can be found in Table A.1.

Cognitive Assessments Programme A collects data from two cognitive assessments taken by applicants. The first is based on the International Cognitive Assessment Resource (ICAR) [34, 165], and has, in various selection cycles, incorporated four different item types: Cube Rotation, Number Sequence, Matrix Reasoning, and Verbal Reasoning. Applicants are given nonverbal and verbal sub-scores, which use a Bayesian generalized linear item response model [28]. In some cycles, only the nonverbal score was used, while other cycles combined the two to create one singular score. The second cognitive assessment is a gameified skills test; methodological details have been redacted to preserve programme anonymity.

Peer Review Stage one applicant essays were judged by two types of human evaluators: other applicants (peers) and adults with some expertise on the project topics (experts). Though Anvari et al. [6] provide evidence for the efficiency and effectiveness of peer review as a measurement of aptitude, peer review was (and remains) experimental [146]. That said, Schee et al. [155] find that decision subjects of a blind peer review process experience just outcomes both according to the similar treatment and similar outcomes principles. Thus, though Programme A treated peer reviews as experimental, peer scoring played an integral role in the Programme A process. To collect peer reviews, each applicant was assigned to review a fixed number of essays submitted by other applicants. Each review consisted

of Likert-scale judgements designed to measure several cognitive, character, and motivational traits relevant to the programme's selection criteria; the specific trait list has been generalised to preserve programme anonymity.

Expert Review Experts, on the other hand, were only asked to assess applicant project essays. Each reviewer was assigned several projects proportional to their capacity to review. Like peers, experts were asked to review different elements of the project, using Likert scales to gauge how effective the project was at accomplishing what the applicant intended and how impressive the project was relative to other projects in this field.

Finalist Activity Assessment The stage two activities were assessed by selectors through a mix of qualitative and quantitative measures. Each activity was scored on a rubric, and the scores were aggregated to create a final score for each finalist on each activity type. Additionally, selectors were given an option to provide specific qualitative feedback on applicants.

A.3 Programme B

A.3.1 Programme Overview

Programme B is a global scholarship programme that supports talented individuals undertaking university study oriented toward a set of research areas the programme has identified as priority. Specific mission language and operational specifics have been generalised to preserve programme anonymity. At the time of this writing, the programme is in the early stages of operating its selection process; some elements described below describe the programme's planned selection process rather than historical practice.

A.3.2 The Selection Process

The programme's selection process places special emphasis on suitability for its priority research areas. The programme has applicants declare their research

interests of choice and assesses applicants relative to those interests. Diversity-like considerations require the programme to ensure that scholars are chosen across each priority area; however, the programme's theory of change does not lend itself to explicit demographic-diversity considerations (unlike Programme A).

Programme B employs a three-stage selection process. In stage one, applicants submit various application materials asynchronously; the programme selects semi-finalists based on the quality of those materials and the programme's cohort composition goals. In stage two, semi-finalists apply to a partner university, and the university handles its internal selection process; programme applicants who receive university scholarships are dubbed Finalists. In stage three, finalists engage in a series of synchronous activities before final decisions are made by the programme's board.

In stage one of selection, applicants submit their demographic information; selections for research area, course, and preferred project; their education record; a list of achievements; and four written essays speaking to their suitability for the programme. After submitting this application, all applicants are invited to take a cognitive assessment assessing convergent and divergent reasoning.

In stage two, semi-finalists apply to the partner university. In stage three, finalists engage in synchronous activities before final decisions are made by the programme's board. Operational details of stages two and three are not yet finalised at the time of writing and have been omitted; some elements have been generalised to preserve programme anonymity.

A.3.3 Data Collection

Programme B collects and constructs several different aptitude measurements of applicants. This is primarily traditional merit-based measures, e.g., cognitive tests, written essays, or academic transcripts. Additionally, the programme constructs experimental measures from gathered data. We discuss methodologically relevant measures here. A comparison of the measurement categories used by Programme A and Programme B can be found in Table A.1.

Cognitive Assessments Much like Programme A, Programme B uses a cognitive assessment based on the International Cognitive Assessment Resource (ICAR) [34, 165]. Though the details of implementation differ, both programmes use the same four item types and the same scoring algorithm [28]. Additionally, Programme B relies on a divergent-thinking assessment; methodological details have been redacted to preserve programme anonymity. Programme B combines both cognitive scores to compute an overall cognitive assessment score.

AI-driven Assessment of Essays Programme B employs an AI-based scoring method as a preliminary screen on applicants' four written essays. The programme requested that specific methodological details of the implementation not be shared. After these four essays are scored, an overall AI-driven score of applicants is calculated.

Expert Assessment of Applications Stage one applicants whose test scores or AI-driven essay scores merited further consideration were judged by expert human evaluators in two types of reviews: anonymous reviews (where reviewers only had access to applicant essays, grades and achievements) and contextual reviews (where reviewers had access to supporting information such as references or applicant demographics). As compared to Programme A's experts, Programme B engaged adult reviewers in a rigorous training process before qualifying them as expert reviewers.

In each review, experts judged applicants on axes related to specific programme goals (e.g., whether the applicant demonstrated an interest in their chosen research area). Anonymous and contextual reviews were ultimately pooled, and an overall review score was calculated.

Semifinalist and Finalist Assessment Operational details of semi-finalist and finalist assessment are not yet finalised at the time of writing and have been omitted from this thesis.

B
Study Protocols

Contents

B.1 Design Workshops from Chapter 4

We split our N=8 participants into two groups of 4 (G1 and G2) to run two participatory design workshops. As these are group discussions, actual programming deviates from the protocol slightly.

Our research question for both workshops is: "Are SHAP explanations useful?"; however, to frame each workshop, we told both groups that we were interested in the answer to two questions: "What does this technology tell us about the algorithm's scores?" and "How do we envision this technology being used in future selection processes?".

Following this, we gave both groups a 15-minute demonstration of the technology, where we described a sample case, gave some example insights, and answered any questions participants had.

The main task for our workshops consists of hands-on cases examining a SHAP-based "waterfall plot" explanation of a programme applicant's score. An example case can be seen in Figure 4.4. Each case is presented to participants as a single slide on a presented slideshow, with additional questions asked by the researchers to prompt discussion. We show each group 5 different cases and spend an average of 10 minutes on each case.

For each case, we ask some of the following questions to prompt discussion:

1. Why are we viewing this applicant?
2. What comments are we responding to?
3. What does the technology appear to say about this candidate?
4. Does the technology address the comment we are responding to?
5. What does this case say about the algorithm as a whole?
6. Does this case necessitate changes to the algorithm?

Finally, after all cases had been examined, we moved to a short reflection on the technology as a whole. We asked participants to answer:

1. What did you think was useful about the technology presented?
2. What was lacking?
3. How might this be improved?

B.2 Interviews from Chapter 6

For the individual interviews, we follow a semi-structured protocol. Following the methodology of Braun and Clarke [23], we do not limit our analysis to these questions. Instead, we deviate from this script as guided by the conversations and our overarching research questions, then we allow themes to emerge naturally from the data. Our interview research questions (also found in Section 6.4.1) are:

1. What is diversity?

2. Which elements of diversity matter in a selection context? Why?

3. How could technology assist in operationalising diversity?

We interviewed 15 individuals from two different talent identification organisations. We conduct each interview separately. We first ask a few questions about the factors that go into decision-making:

1. We're going to take a step back and discuss a hypothetical selection scenario for a fellowship for a group of young people. In this scenario, you have full control over who is selected.

2. Could you please list some things you think are important in deciding who to accept?

3. (Can skip) Which of (these things) are about the individual applicant's performance?

4. What are (these remaining things) about?

5. (Or:) Why are (these things) important?

We then ask participants to define diversity, to break it down into elements, and to discuss why diversity is important:

1. Now I want to talk about diversity. Keeping your list in mind, can you please define diversity?

2. (If the definition is too short) Could you please elaborate on (pick apart)

3. Why do we care about this definition of diversity?

4. Now, if you were going to break your definition into some elements or facets, what would those be?

5. (If they talk about holistic diversity) What considerations are important when looking at diversity holistically?

6. (If elements are vague) How does (pick a metric) factor into your understanding of (element)?

7. Which elements or considerations are most important?

8. How do we measure these facets of diversity?

9. (If this measurement isn't concrete) Imagine you had a "magic metric" that perfectly measured diversity. What does this metric do?

Next, we run two short exercises from the participatory design literature. The first is called "crazy 8s", wherein participants are given 8 minutes to come up with 8 ideas. For these ideas, we ask participants to think about technologies that might help them better understand diversity in selection:

1. Now I want to talk about technology we can build to support thinking about tradeoffs around diversity. Remember that we're stepping away from existing processes and solutions.

2. We're going to start with an exercise called "crazy 8s". For the next 8 minutes, we're going to spend one minute each developing a technology that might help us better understand diversity in selection. These technologies don't have to make sense or be possible; I just want you to think of things that might help you think through diversity. This activity is difficult; don't worry if you find yourself struggling or sounding silly.

3. Take a second to think about a technology. When you're ready, please describe it.

4. (If the technology is unclear) Could you please elaborate on (unclear part)?

The second is called "the magic app", wherein participants are asked to elaborate on a single idea for an application, waving away technical details as 'magic':

1. Now let's dig deeper into one hypothetical "magic app" designed to help us better understand tradeoffs around diversity in selection. The magic app can do anything you might want it to in any way you might want. What does your magic app do?
2. (If the app has visuals) What do your visuals look like?
3. (If the app is pure text) What sorts of visualisations might help you?
4. (If the app has buttons or sliders) What do your buttons do?
5. (If the app doesn't have any interactivity) How might you interact with this app?
6. Now we are going to split the app out into different "pages"
7. (If they haven't already done this) The individual-level page: each applicant will have their individual-level page, which will say things about that applicant
8. (If they haven't already done this) The cohort-level page: each possible cohort will have its cohort-level page, which will update any time we make changes to the cohort.
9. (For each page) What happens on this page?
10. What is the experience of using this page like?
11. (If the page has visuals) What do your visuals look like?
12. (If the page is pure text) What sorts of visualisations might help you?
13. (If the page has buttons or sliders) What do your buttons do?
14. (If the page doesn't have any interactivity) How might you interact with this app?
15. (For each different feature of the page) What makes (feature) useful to you?
16. Thank you! Is there anything else you would like to add?

B.3 Design Workshops from Chapter 6

For the participatory design workshops, we split our participants by organisation. As some individuals could not attend the second session, we have one group of 6 and another of 7. As these are much larger group discussions, we deviate further from our protocol.

The task for these workshops consists of hands-on sessions with different technologies. The technologies are designed and mocked up based on the thematic analysis of the interviews. These technologies are presented to participants via Miro, where they are free to interact with and annotate them. Our research questions for this workshop are:

1. What prototypes best promote diversity?
2. What elements of these prototypes facilitate their success?

Or, for each prototype: "How and why does this prototype promote diversity in talent identification?". Again, though we write a list of questions targeted at these questions, we do not limit our analysis to these questions [23]. Instead, we deviate from this script as guided by the conversations and our overarching research questions, then we allow themes to emerge naturally from the data [23]. For each technology prototype shown, we have questions:

1. This prototype describes... Are any of you familiar with this?
2. In what follows, we're going to discuss this prototype. Let's start with: is it easy to read for you? What does it say?
3. What questions do you have upon seeing this prototype? Feel free to write these down.
4. How would you use this prototype in a hypothetical selection procedure?
5. How (else) would this prototype fit into your current selection procedure?

6. How would your current selection procedure make the best use of this prototype? Would the process need to be changed? Do you think this would be beneficial?

Finally, at the end of our workshop, after we have covered all of the prototypes, we ask participants to place a star next to their favourite prototype on the Miro board [56, 61].

C

Mathematics and Computation

Contents

C.1 ChatGPT Code Generation for Chapter 5

C.2 Proofs of Submodularity and Monotonicity for Chapter 7

Theorem 3. *Submodularity is closed under weighted addition.*

Proof.

$$
\begin{aligned}
&\forall Y \subseteq U, X \subseteq Y, x \in U \setminus Y, a \geq 0, b \geq 0 : \\
&\quad F_1(X \cup \{x\}) - F_1(X) \leq F_1(Y \cup \{x\}) - F_1(Y) \\
&\quad \wedge F_2(X \cup \{x\}) - F_2(X) \leq F_2(Y \cup \{x\}) - F_2(Y) \\
\implies & a * F_1(X \cup \{x\}) - a * F_1(X) \leq a * F_1(Y \cup \{x\}) - a * F_1(Y) \\
&\quad \wedge b * F_2(X \cup \{x\}) - b * F_2(X) \leq b * F_2(Y \cup \{x\}) - b * F_2(Y) \\
\implies & a * F_1(X \cup \{x\}) - a * F_1(X) + b * F_2(X \cup \{x\}) - b * F_2(X) \\
&\quad \leq a * F_1(Y \cup \{x\}) - a * F_1(Y) + b * F_2(Y \cup \{x\}) - b * F_2(Y) \\
\implies & a * F_1 + b * F_2(X \cup \{x\}) - a * F_1 + b * F_2(X) \\
&\quad \leq a * F_1 + b * F_2(Y \cup \{x\}) - a * F_1 + b * F_2(Y)
\end{aligned}
$$

□

Theorem 4. *Monotonicity is closed under weighted addition.*

Proof.

$$
\begin{aligned}
\forall Y \subseteq U, X \subseteq Y, a \geq 0, b \geq 0 : & F_1(X) \leq F_1(Y) \wedge F_2(X) \leq F_2(Y) \\
& \implies a * F_1(X) \leq a * F_1(Y) \wedge b * F_2(X) \leq b * F_2(Y) \\
& \implies a * F_1(X) + b * F_2(X) \leq a * F_1(Y) + b * F_2(Y) \\
& \implies a * F_1 + b * F_2(X) \leq a * F_1 + b * F_2(Y)
\end{aligned}
$$

□

Theorem 5. *The class of functions F as defined in Equation 7.7 is submodular and monotone.*

Proof. By construction, F is the weighted sum of P and D. But D is the weighted sum of functions δ_g^{prop} and δ_G^{count}. It is trivial to see that P is submodular and monotone. We have already demonstrated that δ_g^{prop} and δ_G^{count} are submodular and monotone. Thus, by Theorems 3 and 4, F is submodular and additive. □

C.3 Proof that Algorithm 1 Approximates the SPF

Here, we prove that the greedy approximation method introduced in Algorithm 1 for SPF is a $(1 - \frac{1}{e})$-approximation for our standard class of diversity functions subject to a cardinality constraint [125].

Recall that an organisation's preference function $f : 2^X -> \mathbb{R}^+$ maps a cohort (a set of applicants) to the weighted sum of a "diversity score" and a "performance score" (both non-negative, real numbers). Here, our applicant pool X is represented as a set with applicants as its members, and possible cohorts $C \subseteq X$ are subsets of X. We prove Appendix C.2 that f is monotonic ($A \subseteq B -> f(A) \leq f(B)$) and submodular ($A \subseteq B \wedge x \notin B -> f_A(x) \geq f_B(x)$ (here, $f_S(e) = f(S \cup \{e\}) - f(S)$ denote the marginal gain of adding element e to set S). We demonstrate elsewhere that many common understandings of diversity are represented by standard diversity functions.

Theorem 6. *Let $(S_0...S_k)$ be a sequence of sets where S_0 is the empty set and $S_{i>0}$ is defined by following Algorithm 1 with any ι, d, and p. Further, let $O := argmax_S(f(S) : |S| = k)$ be the set of size k that maximizes $f := \iota * d + (1 - \iota) * p$. Then $f(S_k) \geq (1 - \frac{1}{e})f(O)$.*

Proof. By induction. Let $o_1...o_k = O$ be any ordering of the elements of O. Let $s_i := S_i - S_{i-1}$ be the element added to S_{i-1} to form S_i.

By monotonicity, we have $\forall i.f(O) \leq f(O \cup S_i)$. We can then write $f(O \cup S_i) = f(O \cup S_i) - f(S_i) + f(S_i) = \sum_{j=1}^{k}(f(S_i \cup o_1...o_j) - f(S_i \cup o_1...o_{j-1}))$. I.e., $f(O \cup S_i) = f(O \cup S_i) - f(S_i) + f(S_i) = \sum_{j=1}^{k} f_{S_i \cup o_1...o_{j-1}}(o_j)$.

By submodularity, $\forall j \in [1...k].f_{S_i \cup o_1...o_{j-1}}(o_j) \leq f_{s_i}(o_j)$. Thus, $f(O) \leq f(S_i) + k * f_{S_i}(s_{i+1})$.

Since Algorithm 1 guarantees that $\forall e \in X - S_i.f_{S_i}(e) \geq f_{S_i}(e)$, it follows that, at every stage, $f(S_{i+1}) - f(S_i) \geq \frac{1}{k}(f(O) - f(S_i))$.

Then induction yields $f(O) - f(S_k) \leq (1 - \frac{1}{k})^k f(O) \leq \frac{1}{e} f(O)$. □

D
Reference Figures and Tables

Contents

D.1 Sample Explanations from Chapter 4

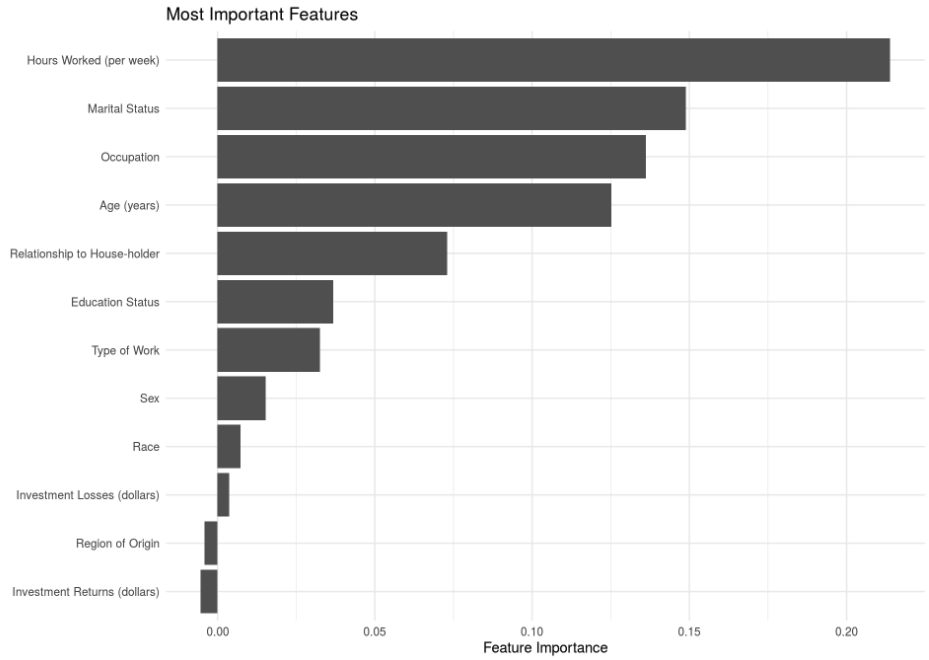


Figure D.1: This figure shows sample SHAP explanations in the *Salary* task. Features can be seen along the y-axis, while feature importance is shown based on the direction and magnitude of the associated bar.

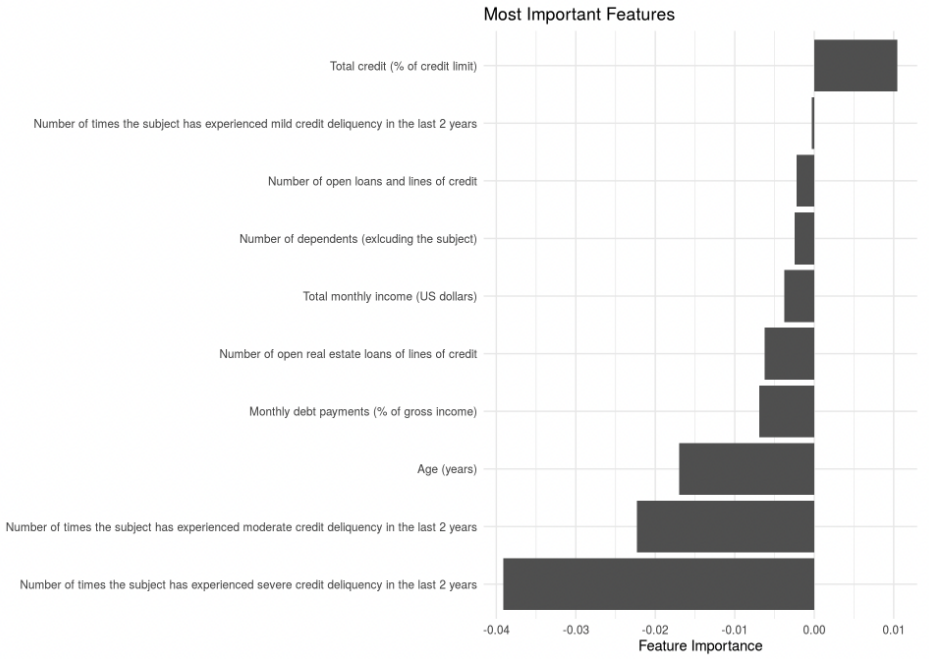


Figure D.2: This figure shows sample SHAP explanations in the *Credit* task. Features can be seen along the y-axis, while feature importance is shown based on the direction and magnitude of the associated bar.

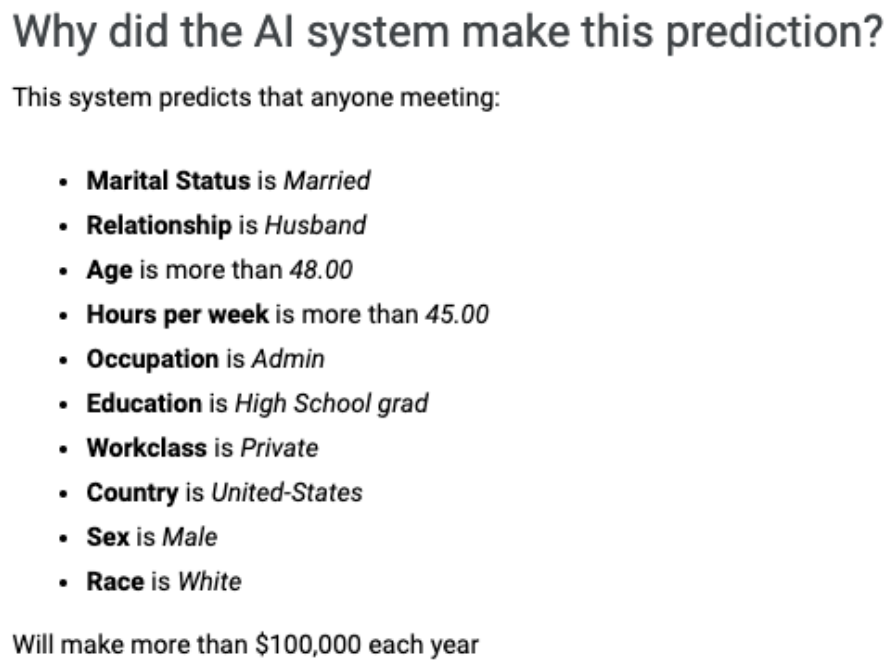


Figure D.3: This figure shows sample Anchor explanations in the *Salary* task. The explanation shows a set of rules that, when jointly followed, increases the likelihood that the model will yield the displayed prediction.

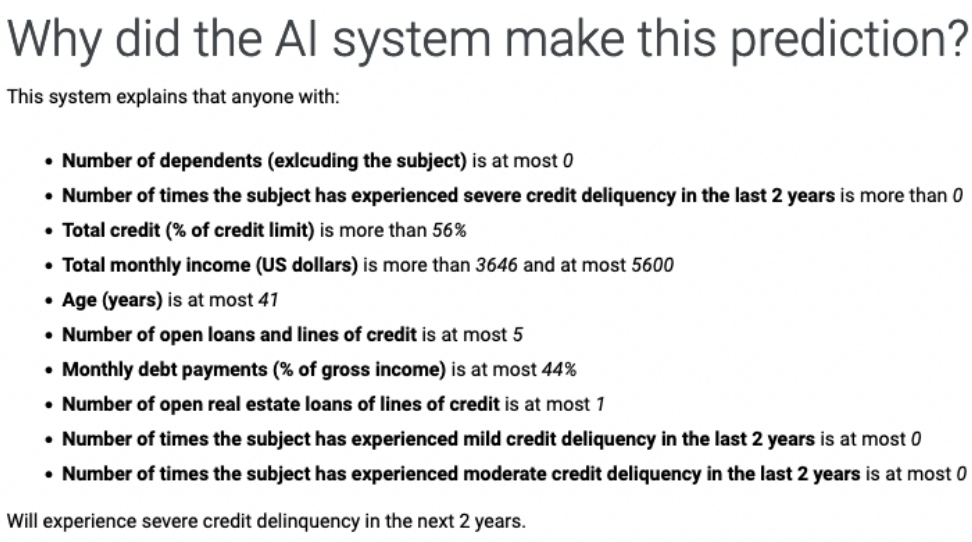


Figure D.4: This figure shows sample Anchor explanations in the *Credit* task. The explanation shows a set of rules that, when jointly followed, increases the likelihood that the model will yield the displayed prediction.

Figure D.5: This figure shows sample Confidence explanations in the *Salary* task. The explanation is simply one sentence containing the model's confidence parameter.

The system is 76% sure of its prediction

Figure D.6: This figure shows sample Confidence explanations in the *Credit* task. The explanation is simply one sentence containing the model's confidence parameter.

D.2 Images and Descriptions of Prototypes from Chapter 6

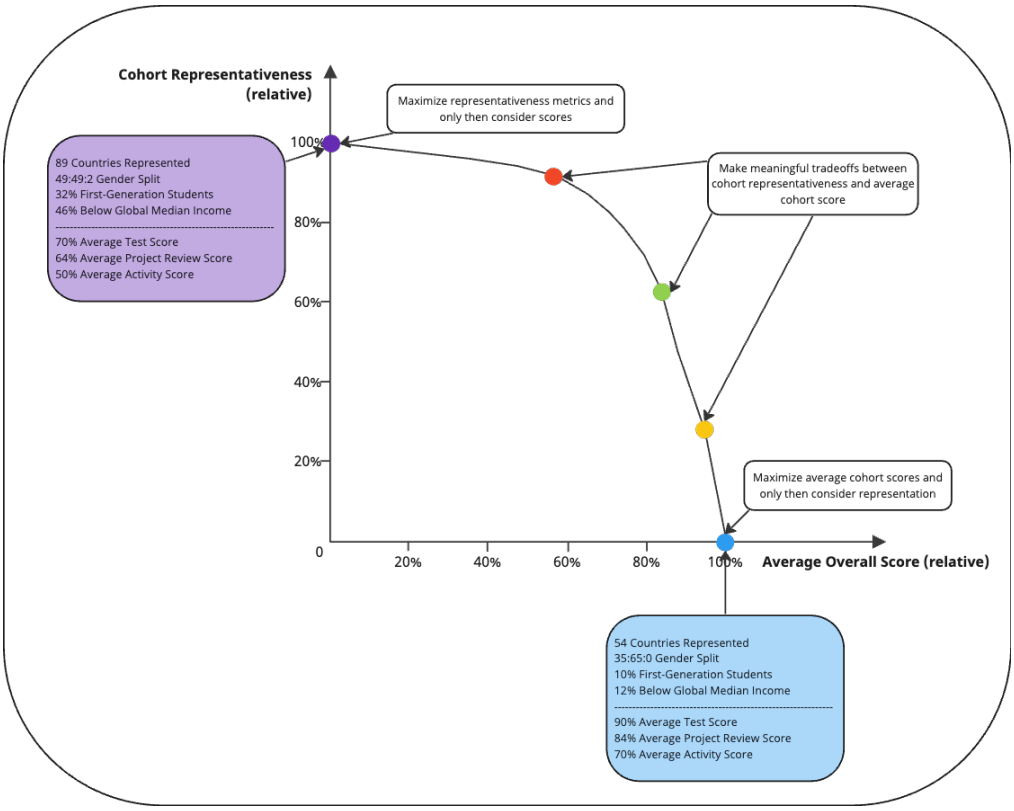


Figure D.7: This figure reproduces Prototype 6.3a at a larger scale.

175

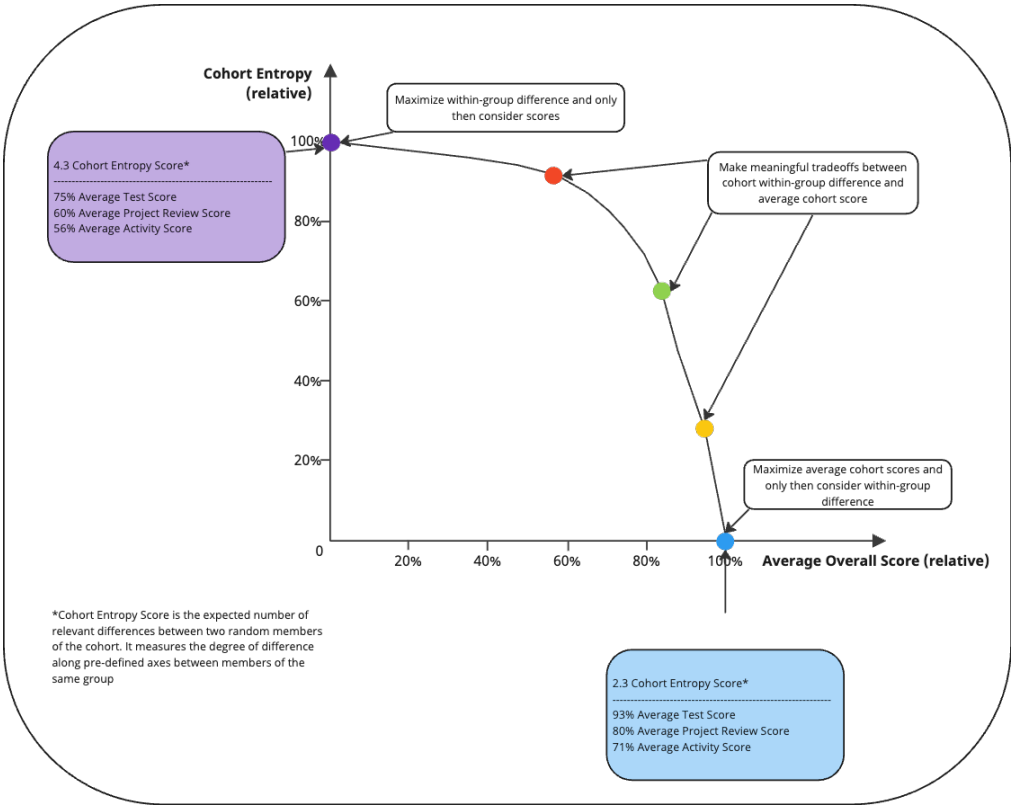


Figure D.8: This figure reproduces Prototype 6.3b at a larger scale.

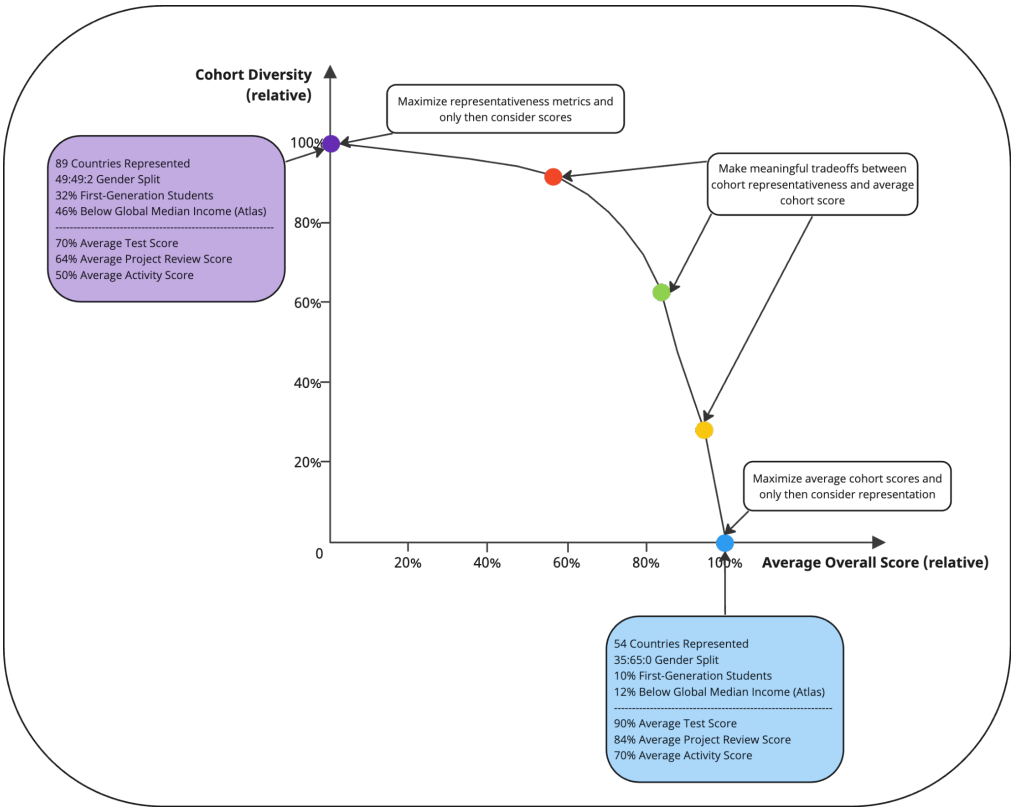


Figure D.9: This figure reproduces Prototype 6.3c at a larger scale.

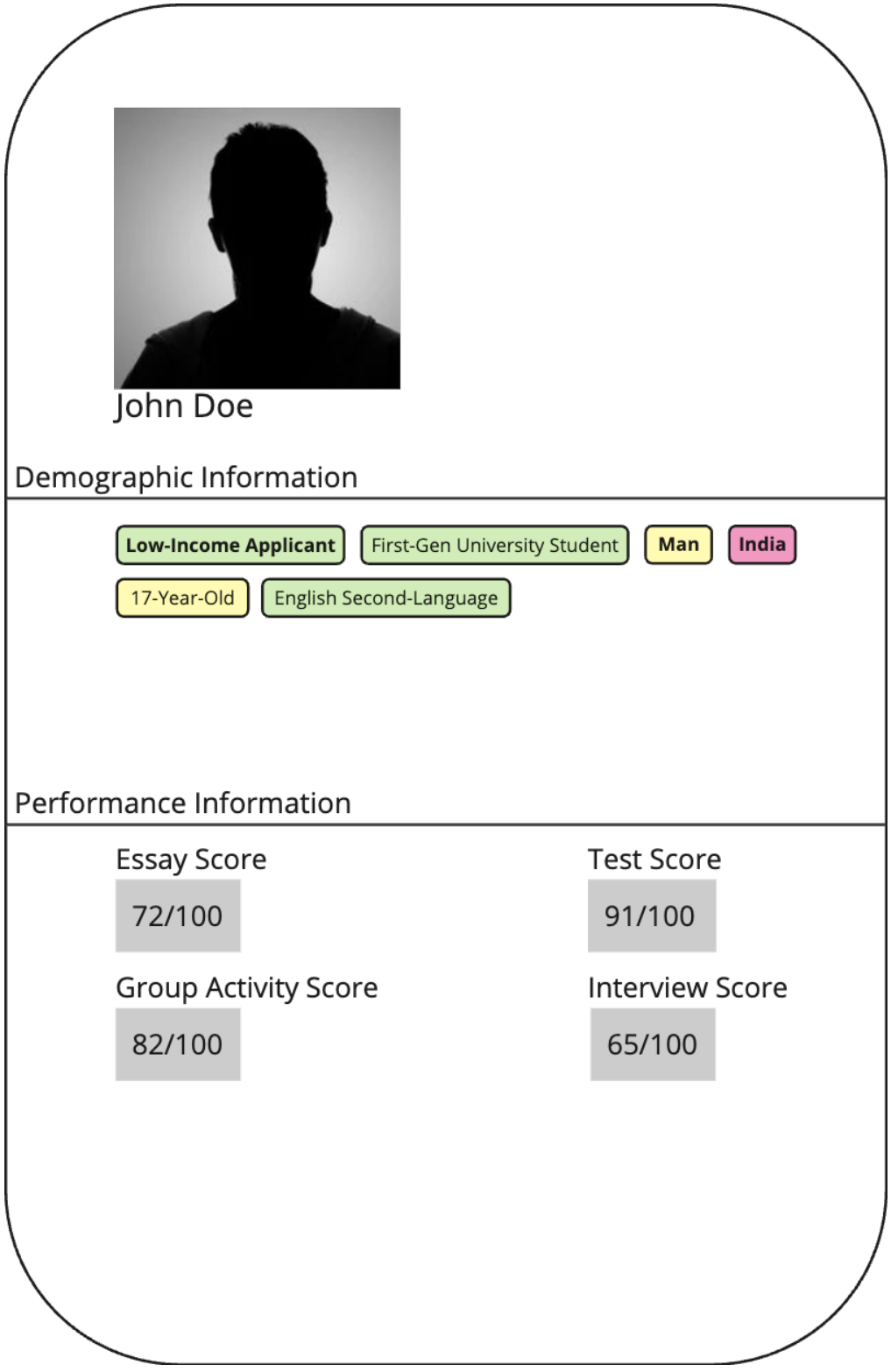


Figure D.10: This figure reproduces Prototype 6.3d at a larger scale.

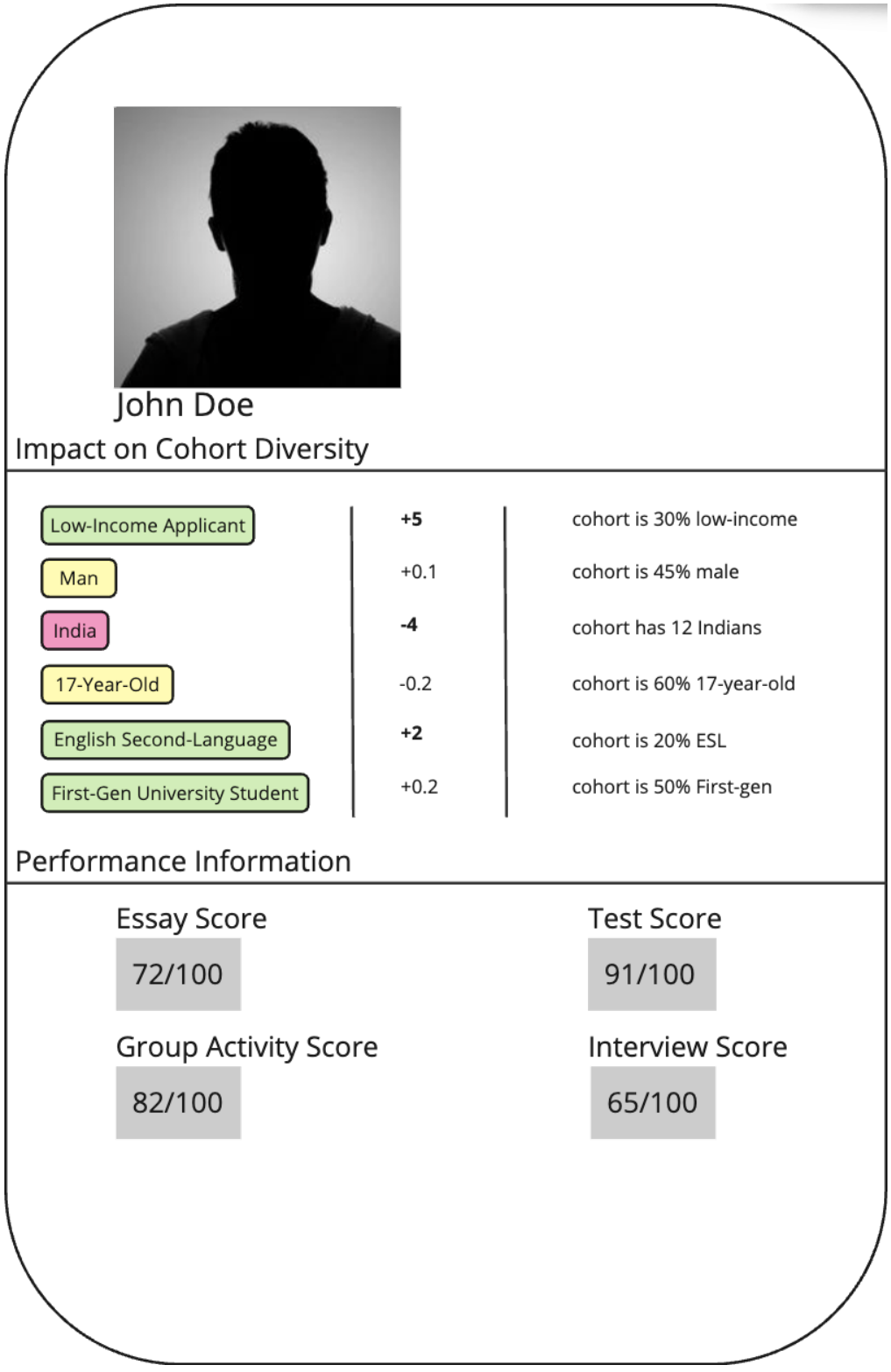


Figure D.11: This figure reproduces Prototype 6.3e at a larger scale.

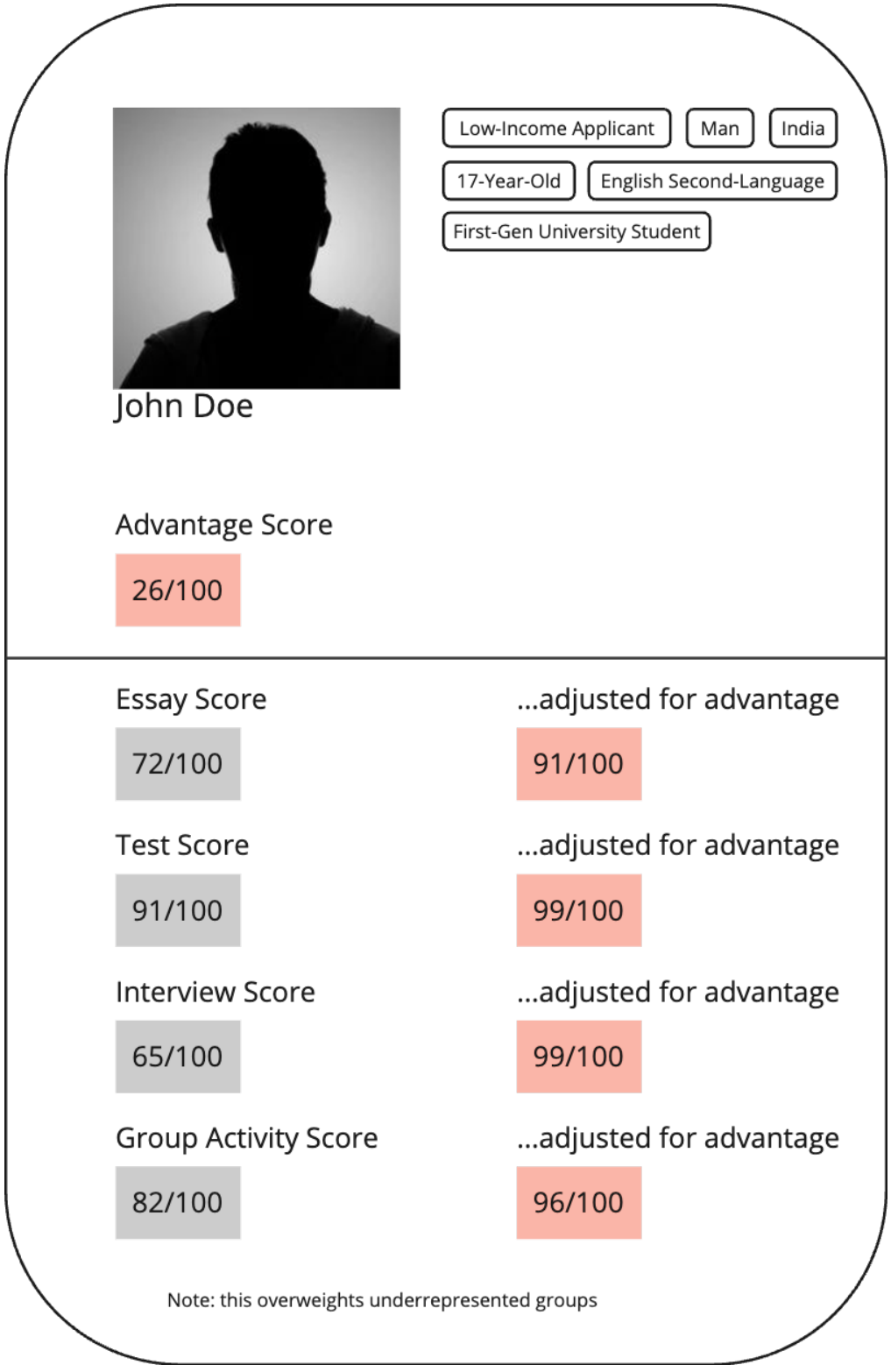


Figure D.12: This figure reproduces Prototype 6.3f at a larger scale.

D.3 Figures and Tables for Chapter 7

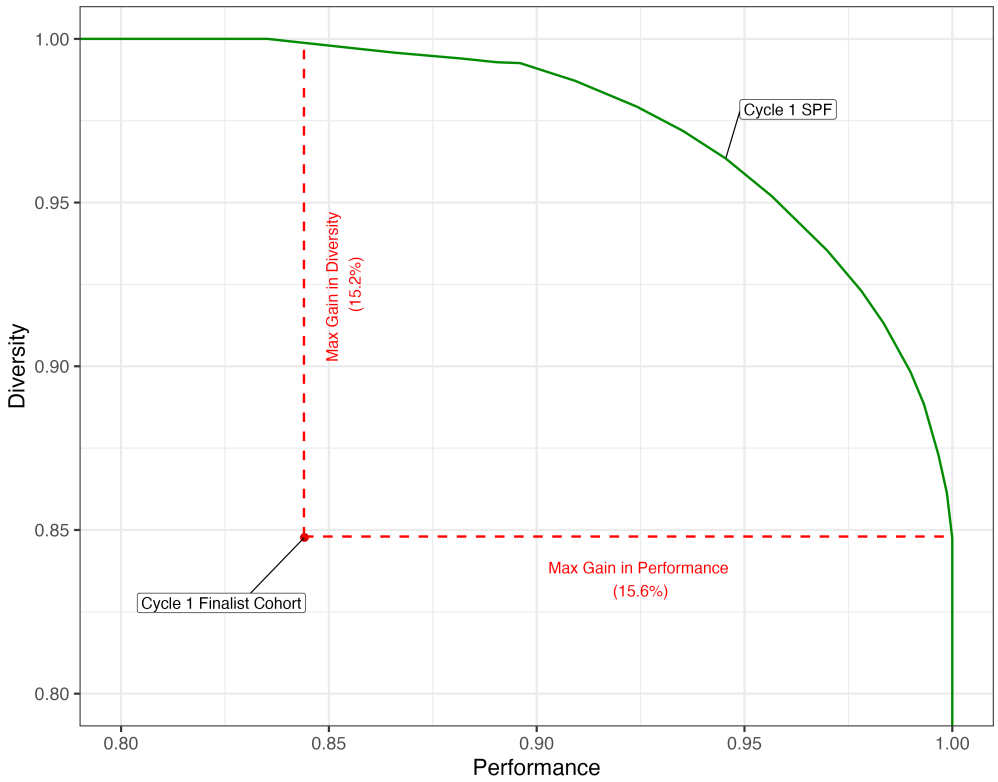


Figure D.13: This figure displays the SPF we estimate for the Cycle W finalist selection process. The y-axis represents the diversity score while the x-axis represents the average cohort performance. The green curve is our estimate of the SPF, which represents the upper bound of diversity that is achievable at every level of cohort performance. The red dot depicts the actual level of diversity and performance of the finalists who were selected in Cycle W. The vertical and horizontal dashed red lines represent the maximum Pareto gain that was possible along the diversity and performance dimensions respectively. In particular, cohort diversity could have been improved by 15.2% without any reduction in cohort performance. And, cohort performance could have been improved by 15.6% without any cost to diversity.

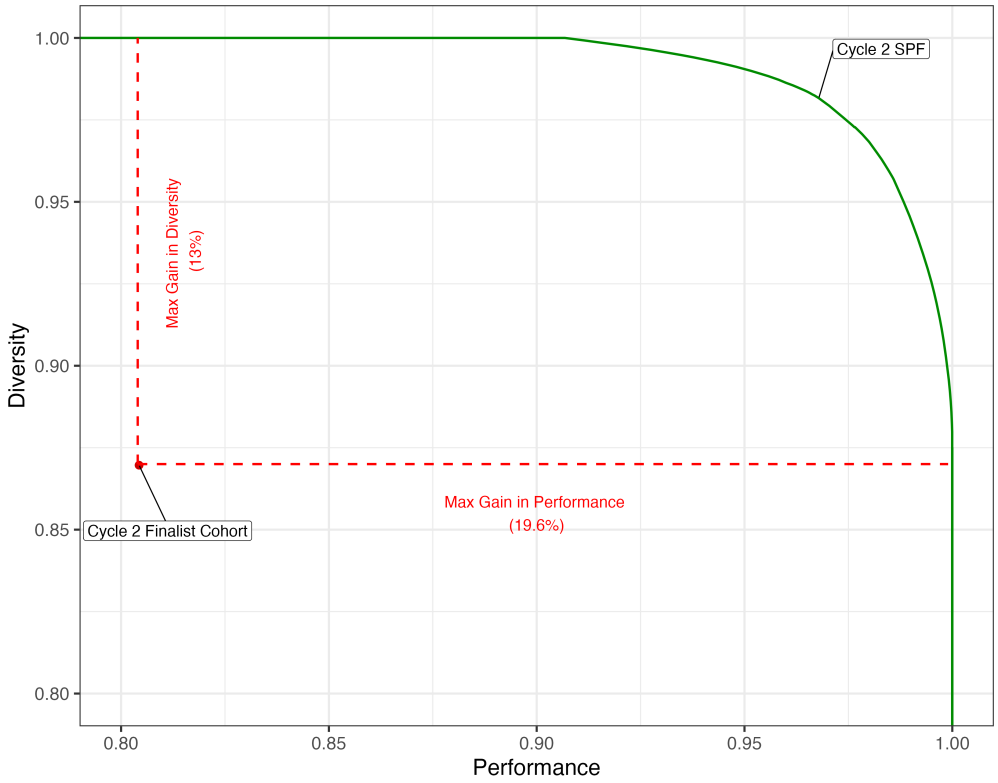


Figure D.14: This figure displays the SPF we estimate for the Cycle X finalist selection process. The y-axis represents the diversity score while the x-axis represents the average cohort performance. The green curve is our estimate of the SPF, which represents the upper bound of diversity that is achievable at every level of cohort performance. The red dot depicts the actual level of diversity and performance of the finalists who were selected in Cycle X. The vertical and horizontal dashed red lines represent the maximum Pareto gain that was possible along the diversity and performance dimensions respectively. In particular, cohort diversity could have been improved by 13% without any reduction in cohort performance. And, cohort performance could have been improved by 19.6% without any cost to diversity.

References

[1] Sara Ahmed. *On Being Included: Racism and Diversity in Institutional Life*. en. Duke University Press, Mar. 2012. URL: https://read.dukeupress.edu/books/book/2209/On-Being-IncludedRacism-and-Diversity-in.

[2] Faiz Ahnaf et al. "AHP and PROMETHEE Comparison on Decision Support System for Scholarship Selection in Universitas Sebelas Maret Surakarta". In: *Proceedings of the International Conference on Industrial Engineering and Operations Management* (2023). URL: https://api.semanticscholar.org/CorpusID:258635921.

[3] Hussam Alkaissi and Samy I McFarlane. "Artificial Hallucinations in ChatGPT: Implications in Scientific Writing". en. In: *Cureus* (Feb. 2023). URL: https://www.cureus.com/articles/138667-artificial-hallucinations-in-chatgpt-implications-in-scientific-writing (visited on 04/08/2023).

[4] Oscar Alvarado and Annika Waern. "Towards Algorithmic Experience: Initial Efforts for Social Media Contexts". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. New York, NY, USA: Association for Computing Machinery, Apr. 2018, pp. 1–12. URL: https://dl.acm.org/doi/10.1145/3173574.3173860.

[5] ELIZABETH ANDERSON. *The Imperative of Integration*. Princeton University Press, 2010. URL: http://www.jstor.org/stable/j.ctt7t225 (visited on 06/11/2025).

[6] Farshid Anvari, Hien Minh Thi Tran, and Deborah Richards. "Effectiveness of Peer Review in Teaching and Learning User Centered Conceptual Design Among Large Cohorts of Information Technology Students". In: *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)* (2021), pp. 66–77. URL: https://api.semanticscholar.org/CorpusID:235639168.

[7] Ruben Arslan, Matthias Walther, and Cyril Tata. "formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R". In: *Behavior Research Methods* 52 (Apr. 2019).

[8] Mallory Avery, Andreas Leibbrandt, and Joseph Vecci. "Does artificial intelligence help or hurt gender diversity? Evidence from two field experiments on recruitment in tech". In: (2024).

[9] Gagan Bansal et al. “Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI ’21. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–16. URL: https://doi.org/10.1145/3411764.3445717 (visited on 01/26/2022).

[10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[11] Solon Barocas, Sophie Hood, and Malte Ziewitz. “Governing Algorithms: A Provocation Piece”. en. In: *SSRN Electronic Journal* (2013). URL: http://www.ssrn.com/abstract=2245322.

[12] Solon Barocas and Andrew D. Selbst. *Big Data’s Disparate Impact*. en. SSRN Scholarly Paper ID 2477899. Rochester, NY: Social Science Research Network, 2016. URL: https://papers.ssrn.com/abstract=2477899 (visited on 10/11/2021).

[13] Solon Barocas, Andrew D. Selbst, and Manish Raghavan. “The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Jan. 2020). arXiv: 1912.04930, pp. 80–89. URL: http://arxiv.org/abs/1912.04930 (visited on 01/05/2022).

[14] Alexander Bastounis et al. *On the consistent reasoning paradox of intelligence and optimal trust in AI: The power of I don’t know*. 2024. URL: https://arxiv.org/abs/2408.02357v1.

[15] Ruha Benjamin. *Race after technology abolitionist tools for the New Jim Code*. eng. Newark: Polity Press, 2019.

[16] Peter Bergman, Elizabeth Kopko, and Julio E Rodriguez. *A Seven-College Experiment Using Algorithms to Track Students: Impacts and Implications for Equity and Fairness*. National Bureau of Economic Research, 2021.

[17] Reuben Binns. *On the Apparent Conflict Between Individual and Group Fairness*. arXiv:1912.06883 [cs, stat]. Dec. 2019. URL: http://arxiv.org/abs/1912.06883 (visited on 04/24/2024).

[18] Reuben Binns et al. “’It’s Reducing a Human Being to a Percentage’; Perceptions of Justice in Algorithmic Decisions”. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Apr. 2018). arXiv: 1801.10408, pp. 1–14. URL: http://arxiv.org/abs/1801.10408 (visited on 01/06/2022).

[19] Zachary Bleemer. “Affirmative action and its race-neutral alternatives”. In: *Journal of Public Economics* 220 (2023), p. 104839.

[20] Mark Blythe. “Research through design fiction: narrative in real and imaginary abstracts”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2014, pp. 703–712.

[21] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. en. The MIT Press, Sept. 1999. URL: https://direct.mit.edu/books/book/4738/Sorting-Things-OutClassification-and-Its.

[22] Hilary Bradbury and Peter Reason. "Action Research: An Opportunity for Revitalizing Research Purpose and Practices". en. In: *Qualitative Social Work* 2.2 (June 2003). Publisher: SAGE Publications, pp. 155–175. URL: https://doi.org/10.1177/1473325003002002003 (visited on 07/30/2024).

[23] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3 (Jan. 2006), pp. 77–101.

[24] Virginia Braun and Victoria Clarke. "Conceptual and design thinking for thematic analysis". In: *Qualitative Psychology* 9.1 (2022). Place: US Publisher: Educational Publishing Foundation, pp. 3–26.

[25] Virginia Braun and Victoria Clarke. "Toward good practice in thematic analysis: Avoiding common problems and be(com)ing a knowing researcher". In: *International Journal of Transgender Health* 24.1 (Jan. 2023). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/26895269.2022.2129597, pp. 1–6. URL: https://doi.org/10.1080/26895269.2022.2129597 (visited on 11/03/2023).

[26] Tom B. Brown et al. *Language Models are Few-Shot Learners*. arXiv:2005.14165 [cs]. July 2020. URL: http://arxiv.org/abs/2005.14165 (visited on 08/24/2022).

[27] Marion Buchenau and Jane Fulton Suri. "Experience prototyping". In: *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*. DIS '00. New York, NY, USA: Association for Computing Machinery, Aug. 2000, pp. 424–433. URL: https://dl.acm.org/doi/10.1145/347642.347802.

[28] Paul-Christian Bürkner. "Bayesian Item Response Modeling in R with brms and Stan". In: *Journal of Statistical Software* 100.5 (2021), pp. 1–54. URL: https://www.jstatsoft.org/index.php/jss/article/view/v100i05.

[29] Federico Cabitza et al. "Explanations Considered Harmful: The Impact of Misleading Explanations on Accuracy in Hybrid Human-AI Decision Making". en. In: *Explainable Artificial Intelligence*. Ed. by Luca Longo, Sebastian Lapuschkin, and Christin Seifert. Cham: Springer Nature Switzerland, 2024, pp. 255–269.

[30] Aaron R. Caldwell et al. *Power Analysis with Superpower*. 2022. URL: https://aaroncaldwell.us/SuperpowerBook/index.html#preface (visited on 09/13/2022).

[31] Le Chen et al. "Investigating the impact of gender on rank in resume search engines". In: *Proceedings of the 2018 chi conference on human factors in computing systems*. 2018, pp. 1–14.

[32] Jorge Chica-Olmo, Angeles Sánchez, and Fabio H. Sepúlveda-Murillo. "Assessing Colombia's policy of socio-economic stratification: An intra-city study of self-reported quality of life". In: *Cities* 97 (2020), p. 102560. URL: https://www.sciencedirect.com/science/article/pii/S0264275119312995.

[33] Danielle Keats Citron. "Technological Due Process". In: *Washington University Law Review* 85.6 (Jan. 2008), pp. 1249–1313.

[34] David M Condon and William Revelle. "The international cognitive ability resource: Development and initial validation of a public-domain measure". In: *Intelligence* 43 (2014), pp. 52–64.

[35] Don Coppersmith and Uzi Vishkin. “Solving NP-hard problems in ‘almost trees’: Vertex cover”. In: *Discrete Applied Mathematics* 10.1 (1985), pp. 27–45. URL: https://www.sciencedirect.com/science/article/pii/0166218X85900575.

[36] Mădălina Cozma, Andrei M. Butnaru, and Radu Tudor Ionescu. *Automated essay scoring with string kernels and word embeddings*. arXiv:1804.07954 [cs]. July 2018. URL: http://arxiv.org/abs/1804.07954 (visited on 08/08/2023).

[37] Will Cukierski Credit Fusion. *Give Me Some Credit*. 2011. URL: https://kaggle.com/competitions/GiveMeSomeCredit.

[38] Kimberle Crenshaw. “Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Policies”. In: *University of Chicago Legal Forum* 1989.1 (1989), pp. 139–167.

[39] Joan R. Dassin, Robin R. Marsh, and Matt Mawer. *International Scholarships in Higher Education*. en. Cham: Springer International Publishing, 2018. URL: http://link.springer.com/10.1007/978-3-319-62734-2.

[40] Clarisse Sieckenius De Souza and Carla Faria Leitão. *Semiotic Engineering Methods for Scientific Research in HCI*. en. Synthesis Lectures on Human-Centered Informatics. Cham: Springer International Publishing, 2009. URL: https://link.springer.com/10.1007/978-3-031-02185-5.

[41] N Dehouche. “Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3)”. en. In: *Ethics in Science and Environmental Politics* 21 (Mar. 2021), pp. 17–23. URL: https://www.int-res.com/abstracts/esep/v21/p17-23/ (visited on 02/20/2024).

[42] David J Deming. “The growing importance of social skills in the labor market”. In: *The quarterly journal of economics* 132.4 (2017), pp. 1593–1640.

[43] Lynn Dombrowski, Ellie Harmon, and Sarah Fox. “Social Justice-Oriented Interaction Design: Outlining Key Design Strategies and Commitments”. en. In: *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. Brisbane QLD Australia: ACM, June 2016, pp. 656–671. URL: https://dl.acm.org/doi/10.1145/2901790.2901861 (visited on 07/30/2024).

[44] Finale Doshi-Velez and Been Kim. “Towards A Rigorous Science of Interpretable Machine Learning”. In: *arXiv:1702.08608 [cs, stat]* (Mar. 2017). arXiv: 1702.08608. URL: http://arxiv.org/abs/1702.08608 (visited on 01/17/2022).

[45] Liam Dugan et al. *RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors*. arXiv:2405.07940 [cs]. June 2024. URL: http://arxiv.org/abs/2405.07940 (visited on 08/01/2024).

[46] Cynthia Dwork et al. “Fairness through awareness”. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS ’12. New York, NY, USA: Association for Computing Machinery, Jan. 2012, pp. 214–226. URL: https://doi.org/10.1145/2090236.2090255 (visited on 01/27/2022).

[47] Mary T. Dzindolet et al. “The role of trust in automation reliance”. en. In: *International Journal of Human-Computer Studies*. Trust and Technology 58.6 (June 2003), pp. 697–718. URL: https://www.sciencedirect.com/science/article/pii/S1071581903000387 (visited on 01/26/2022).

[48] Virginia Eubanks. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. USA: St. Martin's Press, Inc., 2018.

[49] Moran Feldman, Christopher Harshaw, and Amin Karbasi. "Greed Is Good: Near-Optimal Submodular Maximization via Greedy Optimization". en. In: *Proceedings of the 2017 Conference on Learning Theory*. PMLR, June 2017, pp. 758–784. URL: https://proceedings.mlr.press/v65/feldman17b.html.

[50] Will Fleisher. "What's Fair about Individual Fairness?" en. In: 3819799 (July 2021), p. 12. URL: https://papers.ssrn.com/abstract=3819799 (visited on 04/01/2024).

[51] Courtney Ford, Eoin M. Kenny, and Mark T. Keane. "Play MNIST For Me! User Studies on the Effects of Post-Hoc, Example-Based Explanations & Error Rates on Debugging a Deep Learning, Black-Box Classifier". In: *ArXiv* (2020).

[52] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. "On the (im)possibility of fairness". In: arXiv:1609.07236 (Sept. 2016). arXiv:1609.07236 [cs, stat]. URL: http://arxiv.org/abs/1609.07236 (visited on 05/22/2024).

[53] Batya Friedman et al. "Value Sensitive Design and Information Systems". In: Jan. 2006.

[54] T.L. Friedman. *The World Is Flat: A Brief History of the Twenty-first Century*. Business book summary. Farrar, Straus and Giroux, 2005. URL: https://books.google.com/books?id=g3PbAgAAQBAJ.

[55] Gerhard Friedrich and Markus Zanker. "A Taxonomy for Generating Explanations in Recommender Systems". en. In: *AI Magazine* 32.3 (June 2011). Number: 3, pp. 90–98. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/2365 (visited on 01/05/2022).

[56] Amy W. Gatian. "Is user satisfaction a valid measure of system effectiveness?" In: *Information & Management* 26.3 (Mar. 1994), pp. 119–131.

[57] Edmund Gettier. "Is Justified True Belief Knowledge?" In: *Analysis* 23.6 (1963), pp. 121–123.

[58] Phillip Good. *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media, 2013.

[59] Bryce Goodman and Seth Flaxman. "European Union regulations on algorithmic decision-making and a "right to explanation"". In: *AI Magazine* 38.3 (Sept. 2017). arXiv:1606.08813 [cs, stat], pp. 50–57.

[60] GPTZero. *GPTZero | Technology*. Aug. 2023. URL: https://gptzero.me/ (visited on 08/31/2023).

[61] Jillian R. Griffiths, Frances Johnson, and Richard J. Hartley. "User satisfaction as a measure of system performance". en. In: *Journal of Librarianship and Information Science* 39.3 (Sept. 2007), pp. 142–152.

[62] John A Hanley et al. "Receiver operating characteristic (ROC) methodology: the state of the art". In: *Crit Rev Diagn Imaging* 29.3 (1989), pp. 307–335.

[63] X Du-Harpur et al. "What is AI? Applications of artificial intelligence to dermatology". In: *British Journal of Dermatology* 183.3 (2020), pp. 423–430.

[64] John A. Hartigan and Alexandra K. Wigdor, eds. *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery.* Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery. Pages: xii, 354. Washington, DC, US: National Academy Press, 1989.

[65] Gillian R. Hayes. "The relationship of action research to human-computer interaction". In: *ACM Trans. Comput.-Hum. Interact.* 18.3 (Aug. 2011), 15:1–15:20.

[66] Hildebrandt. *Law for computer scientists and other folk.*

[67] Daniel Hirschman, Ellen Berrey, and Fiona Rose-Greenland. "Dequantifying diversity: affirmative action and admissions at the University of Michigan". en. In: *Theory and Society* 45.3 (June 2016), pp. 265–301. URL: https://doi.org/10.1007/s11186-016-9270-2 (visited on 07/10/2024).

[68] Lu Hong and Scott Page. "Groups of Diverse Problem Solvers Can Outperform Groups of High-Ability Problem Solvers". In: *Proceedings of the National Academy of Sciences of the United States of America* 101 (Dec. 2004), pp. 16385–9.

[69] Piotr Horodyski. "Applicants' perception of artificial intelligence in the recruitment process". en. In: *Computers in Human Behavior Reports* 11 (Aug. 2023), p. 100303. URL: https://www.sciencedirect.com/science/article/pii/S2451958823000362 (visited on 08/01/2023).

[70] Chang-Tai Hsieh et al. "The allocation of talent and us economic growth". In: *Econometrica* 87.5 (2019), pp. 1439–1474.

[71] Daniela Huppenkothen, Brian McFee, and Laura Norén. "Entrofy your cohort: A transparent method for diverse cohort selection". In: *Plos one* 15.7 (July 2020), e0231939. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7384611/ (visited on 04/12/2022).

[72] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: (Oct. 2018). MAG ID: 2963341956, pp. 4171–4186.

[73] Maia Jacobs et al. "How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection". en. In: *Translational Psychiatry* 11.1 (Feb. 2021). Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Depression;Scientific community Subject_term_id: depression;scientific-community, pp. 1–9. URL: https://www.nature.com/articles/s41398-021-01224-x (visited on 01/26/2022).

[74] Sheila Jasanoff. *States of Knowledge.* en. 0th ed. Routledge, July 2004. URL: https://www.taylorfrancis.com/books/9781134328345.

[75] Dilrabo Jonbekova et al. "How international higher education graduates contribute to their home country: an example from government scholarship recipients in Kazakhstan". In: *Higher Education Research and Development* 42.1 (2023), pp. 126–140.

[76] Justin Kaashoek, Manish Raghavan, and John J. Horton. "The Impact of Generative AI on Labor Market Matching". In: *An MIT Exploration of Generative AI* (2024). https://mit-genai.pubpub.org/pub/4t8pqt06.

[77] Kalpesh Krishna et al. "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense". In: (2023). ARXIV_ID: 2303.13408 S2ID: 2969e8a14237f8244d3c825ff19bdfb3cc7fddf1.

[78] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. "Algorithmic Recourse: from Counterfactual Explanations to Interventions". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 353–362. URL: https://doi.org/10.1145/3442188.3445899 (visited on 10/15/2021).

[79] Michael Kearns et al. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2564–2572. URL: https://proceedings.mlr.press/v80/kearns18a.html.

[80] Hae-Young Kim. "Statistical notes for clinical researchers: post-hoc multiple comparisons". In: *Restorative Dentistry and Endodontics* 40.2 (May 2015), pp. 172–176.

[81] Jan Hendrik Kirchner et al. *New AI classifier for indicating AI-written text*. en-US. Jan. 2023. URL: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text (visited on 08/31/2023).

[82] Jon Kleinberg et al. "Algorithmic fairness". In: *Aea papers and proceedings*. Vol. 108. 2018, pp. 22–27.

[83] Jake Knapp, John Zeratzky, and Braden Kowitz. *Sprint: How to solve big problems and test new ideas in just five days*. Simon and Schuster, Mar. 2016.

[84] Ron Kohavi. "Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, Aug. 1996, pp. 202–207. (Visited on 09/12/2022).

[85] Andreas Krause and Daniel Golovin. "Submodular Function Maximization". en. In: *Tractability*. Ed. by Lucas Bordeaux, Youssef Hamadi, and Pushmeet Kohli. 1st ed. Cambridge University Press, Feb. 2014, pp. 71–104. URL: https://www.cambridge.org/core/product/identifier/CBO9781139177801A031/type/book_part (visited on 02/15/2023).

[86] Andreas Krause and Daniel Golovin. "Submodular function maximization." In: *Tractability* 3.71-104 (2014), p. 3.

[87] Satyapriya Krishna et al. "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective". en. In: (Feb. 2022). URL: https://arxiv.org/abs/2202.01602v2 (visited on 02/08/2022).

[88] I. Elizabeth Kumar et al. "Problems with Shapley-value-based explanations as feature importance measures". In: *arXiv:2002.11097 [cs, stat]* (June 2020). arXiv: 2002.11097. URL: http://arxiv.org/abs/2002.11097 (visited on 01/05/2022).

[89] Preethi Lahoti, Krishna P. Gummadi, and Gerhard Weikum. *iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making*. 2019. arXiv: 1806.01059 [cs.LG]. URL: https://arxiv.org/abs/1806.01059.

[90] Vivian Lai and Chenhao Tan. "On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT*'19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 29–38. URL: https://doi.org/10.1145/3287560.3287590 (visited on 01/26/2022).

[91] Anja Lambrecht and Catherine Tucker. "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads". In: *Management science* 65.7 (2019), pp. 2966–2981.

[92] Jeff Larson et al. *How We Analyzed the COMPAS Recidivism Algorithm*. en. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (visited on 11/09/2022).

[93] Mitra Lashkari and Jinghui Cheng. ""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. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–16. URL: https://dl.acm.org/doi/10.1145/3544548.3581548.

[94] Michael Latzer et al. "The Economics of Algorithmic Selection on the Internet". en. In: 2710399. Rochester, NY, Oct. 2014. URL: https://papers.ssrn.com/abstract=2710399.

[95] David B. Leake. "Artiicial Intelligence". In: 2001. URL: https://api.semanticscholar.org/CorpusID:18409035.

[96] Weiwen Leung et al. "Race, Gender and Beauty: The Effect of Information Provision on Online Hiring Biases". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–11. URL: https://dl.acm.org/doi/10.1145/3313831.3376874.

[97] Patrick Lewis et al. "Retrieval-augmented generation for knowledge-intensive nlp tasks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[98] Danielle Li, Lindsey R Raymond, and Peter Bergman. *Hiring as Exploration*. Tech. rep. National Bureau of Economic Research, 2020.

[99] Weixin Liang et al. "GPT detectors are biased against non-native English writers". In: *Patterns* 4.7 (July 2023), p. 100779. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10382961/ (visited on 09/20/2023).

[100] Zachary C Lipton. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3 (2018), pp. 31–57.

[101] Zhengliang Liu et al. *DeID-GPT: Zero-shot Medical Text De-Identification by GPT-4*. arXiv:2303.11032 [cs]. Mar. 2023. URL: http://arxiv.org/abs/2303.11032 (visited on 04/05/2023).

[102] Alex Jiahong Lu et al. "Organizing Community-based Events in Participatory Action Research: Lessons Learned from a Photovoice Exhibition". In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI EA '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–8. URL: https://dl.acm.org/doi/10.1145/3544549.3573846 (visited on 07/30/2024).

[103] Scott Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *arXiv:1705.07874 [cs, stat]* (Nov. 2017). arXiv: 1705.07874. URL: http://arxiv.org/abs/1705.07874 (visited on 01/06/2022).

[104] Alexandra Mai and Katharina Pfeffer. "User Mental Models of Cryptocurrency Systems - A Grounded Theory Approach". en. In: ().

[105] Frank Marcinkowski et al. "Implications of AI (un-)fairness in higher education admissions: the effects of perceived AI (un-)fairness on exit, voice and organisational reputation". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT*'20. Barcelona, Spain: Association for Computing Machinery, 2020, pp. 122–130. URL: https://doi.org/10.1145/3351095.3372867.

[106] Elijah Mayfield. "Defensible Explanations for Algorithmic Decsions about Writing in Education". PhD thesis. Carnegie Mellon University, 2020.

[107] Ninareh Mehrabi et al. "A Survey on Bias and Fairness in Machine Learning". In: *ACM Comput. Surv.* 54.6 (July 2021). URL: https://doi.org/10.1145/3457607.

[108] Craig A Mertler. *The Wiley handbook of action research in education*. en. Wiley handbooks in education. Hoboken, NJ: John Wiley and Sons, 2019.

[109] John Stuart Mill. *On Liberty*. Cambridge Library Collection - Philosophy. Cambridge University Press, 2011.

[110] Tim Miller. "Explanation in Artificial Intelligence: Insights from the Social Sciences". en. In: *CoRR* abs/1706.07269 (June 2017). _eprint: 1706.07269. URL: https://arxiv.org/abs/1706.07269v3 (visited on 11/08/2021).

[111] Tim Miller. *Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven decision support*. arXiv:2302.12389 [cs]. Mar. 2023. URL: http://arxiv.org/abs/2302.12389 (visited on 04/05/2023).

[112] Charles W Mills. *Blackness visible: Essays on philosophy and race*. Cornell University Press, 2015.

[113] Rachel Minkin. "Diversity, Equity and Inclusion in the Workplace". In: (2023).

[114] Prabhaker Mishra et al. "Application of Student's t-test, Analysis of Variance, and Covariance". In: *Annals of Cardiac Anaesthesia* 22.4 (2019), pp. 407–411.

[115] Eric Mitchell et al. *DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature*. arXiv:2301.11305 [cs]. Jan. 2023. URL: http://arxiv.org/abs/2301.11305 (visited on 04/06/2023).

[116] Margaret Mitchell et al. “Diversity and Inclusion Metrics in Subset Selection”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 117–123. URL: https://doi.org/10.1145/3375627.3375832.

[117] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2019.

[118] María de los Ángeles Morata-Ramírez and Francisco Pablo Holgado-Tello. “Construct Validity of Likert Scales through Confirmatory Factor Analysis: A Simulation Study Comparing Different Methods of Estimation Based on Pearson and Polychoric Correlations”. en. In: *International Journal of Social Science Studies* 1.1 (Jan. 2013). Number: 1, pp. 54–61. URL: https://redfame.com/journal/index.php/ijsss/article/view/27 (visited on 09/15/2022).

[119] Aldon D Morris. *The origins of the civil rights movement*. Simon and Schuster, 1984.

[120] Kathleen L. Mosier et al. “Automation Bias, Accountability, and Verification Behaviors”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40.4 (1996), pp. 204–208. eprint: https://doi.org/10.1177/154193129604000413. URL: https://doi.org/10.1177/154193129604000413.

[121] Neil Natarajan. “Human-AI Collaboration in Recruitment and Selection”. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. Ed. by Edith Elkind. Doctoral Consortium. International Joint Conferences on Artificial Intelligence Organisation, Aug. 2023, pp. 7089–7090. URL: https://doi.org/10.24963/ijcai.2023/819.

[122] Neil Natarajan and Reuben Binns. *Misleading AI Explanations*. Aug. 2022. URL: osf.io/mq86p.

[123] Neil Natarajan, Elías Sánchez Hanno, and Logan Gittelson. “Detecting Generative AI Usage in Application Essays”. en. In: *Generative AI and HCI workshop at CHI 2024*. 2024. URL: https://generativeaiandhci.github.io/papers/2024/genaichi2024_9.pdf.

[124] Neil Natarajan et al. “Trust Explanations to Do What They Say”. en. In: *Human-Centered AI Workshop at NeurIPS 2022*. 2022. URL: https://openreview.net/pdf?id=mzsPCefDaY5.

[125] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. “An analysis of approximations for maximizing submodular set functions—I”. In: *Mathematical programming* 14 (1978), pp. 265–294.

[126] Stella M Nkomo et al. “Diversity at a critical juncture: New theories for a complex phenomenon”. In: *Academy of Management Review* 44.3 (2019), pp. 498–517.

[127] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018. URL: http://www.jstor.org/stable/j.ctt1pwt9w5 (visited on 06/11/2025).

[128] Kadeem Noray and Savannah Noray. *Communication and Systemic Disadvantage: Evidence from the Increase in Social Skills*. Tech. rep. Harvard University, 2023.

[129] Serena Olsaretti. *The Oxford Handbook of Distributive Justice*. Oxford University Press, May 2018. URL: https://doi.org/10.1093/oxfordhb/9780199645121.001.0001.

[130] OpenAI. *GPT-4 Technical Report*. arXiv:2303.08774 [cs]. Mar. 2023. URL: http://arxiv.org/abs/2303.08774 (visited on 08/31/2023).

[131] Peter Organisciak et al. "Beyond semantic distance: Automated scoring of divergent thinking greatly improves with large language models". In: *Thinking Skills and Creativity* 49 (Sept. 2023), p. 101356. URL: https://www.sciencedirect.com/science/article/pii/S1871187123001256 (visited on 03/04/2024).

[132] Scott Page. *The difference: How the power of diversity creates better groups, firms, schools, and societies-new edition*. Princeton University Press, 2008.

[133] Scott E. Page. "Diversity and Complexity". In: (Nov. 2010). MAG ID: 1653902796.

[134] Scott E. Page, Earl Lewis, and Nancy Cantor. "The Diversity Bonus: How Great Teams Pay Off in the Knowledge Economy". In: (Sept. 2017). MAG ID: 2885284796 S2ID: e47d67055846c132fc8274170fe9adf1250a7727.

[135] Suruchi Pandey and Medha Bahukhandi. "Applicants' Perception Towards the Application of AI in Recruitment Process". In: *2022 Interdisciplinary Research in Technology and Management (IRTM)*. Feb. 2022, pp. 1–6.

[136] Christos H. Papadimitriou. "Computational complexity". In: *Encyclopedia of Computer Science*. GBR: John Wiley and Sons Ltd., 2003, pp. 260–265.

[137] Andrea Papenmeier et al. "It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI". en. In: *ACM Transactions on Computer-Human Interaction* 29.4 (Aug. 2022), pp. 1–33. URL: https://dl.acm.org/doi/10.1145/3495013 (visited on 12/05/2023).

[138] Frank Pasquale. "Rankings, Reductionism, and Responsibility". In: *Faculty Scholarship* (Jan. 2006). URL: https://digitalcommons.law.umaryland.edu/fac_pubs/1351.

[139] Andi Peng et al. "What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7.1 (2019), pp. 125–134. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/5281.

[140] Mike Perkins and Jasper Roe. "Decoding Academic Integrity Policies: A Corpus Linguistics Investigation of AI and Other Technological Threats". In: (July 2023). DOI: 10.1057/s41307-023-00323-2 MAG ID: 4384560074.

[141] Uwe Peters. "Hidden figures: epistemic costs and benefits of detecting (invisible) diversity in science". In: *European Journal for Philosophy of Science* 11.1 (2021), p. 33.

[142] Rajasshrie Pillai and Brijesh Sivathanu. "Adoption of artificial intelligence (AI) for talent acquisition in IT/ITeS organisations". In: *Benchmarking: An International Journal* 27.9 (Aug. 2020). MAG ID: 3049223401, pp. 2599–2629.

[143] Randal Pinkett. *Data-driven DEI: The tools and metrics you need to measure, analyze, and improve diversity, equity, and inclusion*. John Wiley & Sons, 2023.

[144] Martha Poon. “Scorecards as devices for consumer credit: the case of Fair, Isaac & Company Incorporated”. In: *The Sociological Review* 55.s2 (2007), pp. 284–306. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-954X.2007.00740.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-954X.2007.00740.x.

[145] Manish Raghavan et al. “Mitigating bias in algorithmic hiring: Evaluating claims and practices”. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. FAT*’20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 469–481. URL: https://dl.acm.org/doi/10.1145/3351095.3372828 (visited on 08/01/2023).

[146] Refanja Rahmatillah and Rizki Fajrita. “Analyzing Factors Affecting the Effectiveness of Peer Assessment in EFL Teaching”. In: *Indonesian Journal of Teaching and Teacher Education* (2022). URL: https://api.semanticscholar.org/CorpusID:256757919.

[147] Dadi Ramesh and Suresh Kumar Sanampudi. “An automated essay scoring systems: a systematic literature review”. en. In: *Artificial Intelligence Review* 55.3 (Mar. 2022), pp. 2495–2527. URL: https://doi.org/10.1007/s10462-021-10068-2 (visited on 08/11/2023).

[148] Felix G. Rebitschek, Gerd Gigerenzer, and Gert G. Wagner. “People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors”. en. In: *Scientific Reports* 11.1 (2021), p. 20171.

[149] Howard Rheingold. *Smart Mobs: The Next Social Revolution*. Perseus Publishing, 2002.

[150] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. en. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, Aug. 2016, pp. 1135–1144. URL: https://dl.acm.org/doi/10.1145/2939672.2939778 (visited on 01/21/2022).

[151] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance”. In: *arXiv:1611.05817 [cs, stat]* (Nov. 2016). arXiv: 1611.05817. URL: http://arxiv.org/abs/1611.05817 (visited on 01/05/2022).

[152] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Anchors: High Precision Model-Agnostic Explanations”. en. In: *AAAI*. 2018, p. 9.

[153] Michael Roy. “Cathy O’Neil. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy . New York: Crown Publishers, 2016. 272p. Hardcover, $26 (ISBN 978-0553418811).” In: *College & Research Libraries* 78 (Mar. 2017), pp. 403–404.

[154] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1.5 (May 2019), pp. 206–215.

[155] Brian A Vander Schee, Tony Stovall, and Demetra Andrews. "Using cross-course peer grading with content expertise, anonymity, and perceived justice". In: *Active Learning in Higher Education* 25 (2022), pp. 101–114. URL: https://api.semanticscholar.org/CorpusID:248899447.

[156] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. "How computers see gender: An evaluation of gender classification in commercial facial analysis services". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–33.

[157] Frank L Schmidt and John E Hunter. "The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings." In: *Psychological bulletin* 124.2 (Jan. 1998). MAG ID: 2136971664, p. 262.

[158] Patrick Schober, Christa Boer, and Lothar A. Schwarte. "Correlation Coefficients: Appropriate Use and Interpretation". eng. In: *Anesthesia and Analgesia* 126.5 (May 2018), pp. 1763–1768.

[159] Candice Schumann et al. "The diverse cohort selection problem". In: *arXiv preprint arXiv:1709.03441* (Mar. 2017). arXiv: 1709.03441. URL: http://arxiv.org/abs/1709.03441 (visited on 04/12/2022).

[160] Nick Seaver. "Algorithms as culture: Some tactics for the ethnography of algorithmic systems". In: *Big Data & Society* 4.2 (2017), p. 2053951717738104.

[161] Yonadav Shavit et al. "Practices for Governing Agentic AI Systems". en. In: ().

[162] Edward Shortliffe. *Computer-Based Medical Consultations: Mycin*. Vol. 388. Elsevier, Oct. 1976, pp. 243–260. URL: https://www.sciencedirect.com/science/article/pii/B9780444001795500147.

[163] Daniel Steel et al. "Multiple diversity concepts and their ethical-epistemic implications". en. In: *European Journal for Philosophy of Science* 8.3 (Oct. 2018), pp. 761–780. URL: https://doi.org/10.1007/s13194-018-0209-5 (visited on 02/28/2024).

[164] Amber L. Stephenson and David B. Yerger. "Does brand identification transform alumni into university advocates?" en. In: *International Review on Public and Nonprofit Marketing* 11.3 (Oct. 2014), pp. 243–262.

[165] Siniša Subotić et al. "Psychometric validation of the ICAR Matrix Reasoning test". In: *Empirical Studies in Psychology* 59 (2020), p. 62.

[166] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. "Does fair ranking improve minority outcomes? understanding the interplay of human and algorithmic biases in online hiring". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 989–999.

[167] Mo-Yin S Tam, Gilbert W Bassett, and Uday Sukhatme. "New selection indices for university admissions: A quantile approach". en. In: *Statistical Data Analysis Based on the L 1-Norm and Related Methods*. Ed. by Yadolah Dodge. Basel: Birkhäuser Basel, 2002, pp. 67–76. URL: http://link.springer.com/10.1007/978-3-0348-8201-9_6 (visited on 04/12/2022).

[168] Tharindu Kumarage et al. “Stylometric Detection of AI-Generated Text in Twitter Timelines”. In: *ArXiv* (2023). ARXIV_ID: 2303.03697 S2ID: 18e0b11dd5b1b413d33308da0379836752aaaec1.

[169] *Thematic Analysis | SAGE Publications Ltd*. URL: https://uk.sagepub.com/en-gb/eur/thematic-analysis/book248481 (visited on 02/24/2023).

[170] H Holden Thorp. “ChatGPT is fun, but not an author”. In: 379.6630 (Jan. 2023). MAG ID: 4318263917, pp. 313–313.

[171] Berk Ustun, Alexander Spangher, and Yang Liu. “Actionable Recourse in Linear Classification”. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Jan. 2019). arXiv: 1809.06514, pp. 10–19. URL: http://arxiv.org/abs/1809.06514 (visited on 01/05/2022).

[172] Max Van Kleek et al. “Respectful things: Adding social intelligence to “smart” devices”. In: *Living in the Internet of Things: Cybersecurity of the IoT - 2018*. Mar. 2018, pp. 1–6. URL: https://ieeexplore.ieee.org/document/8379693.

[173] Ashish Vaswani et al. “Attention is All you Need”. In: 30 (June 2017). MAG ID: 2963403868, pp. 5998–6008.

[174] Megan Venn-Wycherley et al. “The Realities of Evaluating Educational Technology in School Settings”. In: *ACM Trans. Comput.-Hum. Interact.* 31.2 (Feb. 2024), 26:1–26:33. URL: https://doi.org/10.1145/3635146 (visited on 07/12/2024).

[175] Vivek Verma et al. *Ghostbuster: Detecting Text Ghostwritten by Large Language Models*. arXiv:2305.15047 [cs]. Nov. 2023. URL: http://arxiv.org/abs/2305.15047 (visited on 03/18/2024).

[176] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR”. In: *CoRR* abs/1711.00399 (2017). _eprint: 1711.00399. URL: http://arxiv.org/abs/1711.00399.

[177] Pei Wang. “What Do You Mean by" AI"?” In: *AGI*. Vol. 171. 2008, pp. 362–373.

[178] Yongjie Wang et al. *On the Use of BERT for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation*. arXiv:2205.03835 [cs]. May 2022. URL: http://arxiv.org/abs/2205.03835 (visited on 08/08/2023).

[179] Natasha Warikoo. *The Diversity Bargain: And Other Dilemmas of Race, Admissions, and Meritocracy at Elite Universities*. en. Chicago, IL: University of Chicago Press, Feb. 2019. URL: https://press.uchicago.edu/ucp/books/book/chicago/D/bo24550619.html.

[180] Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. “A Human-Grounded Evaluation of SHAP for Alert Processing”. In: *arXiv:1907.03324 [cs, stat]* (July 2019). arXiv: 1907.03324. URL: http://arxiv.org/abs/1907.03324 (visited on 04/14/2022).

[181] Drew Westen and Robert Rosenthal. “Quantifying construct validity: Two simple measures”. In: *Journal of Personality and Social Psychology* 84 (2003). Place: US Publisher: American Psychological Association, pp. 608–618.

[182] Paris Will, Dario Krpan, and Grace Lordan. "People versus machines: introducing the HIRE framework". In: *Artificial Intelligence Review* 56.2 (2023), pp. 1071–1100.

[183] Alison Wylie. "Introduction: when difference makes a difference". In: *Episteme* 3.1-2 (2006), pp. 1–7.

[184] Lynette Yarger, Fay Cobb Payton, and Bikalpa Neupane. "Algorithmic equity in the hiring of underrepresented IT job candidates". In: *Online information review* 44.2 (2020), pp. 383–395.

[185] Iris Marion Young. *Justice and the Politics of Difference*. Princeton University Press, 1990.

[186] Rich Zemel et al. "Learning Fair Representations". In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research 3. Atlanta, Georgia, USA: PMLR, 2013, pp. 325–333. URL: https://proceedings.mlr.press/v28/zemel13.html.

[187] Yuying Zhao et al. "Fairness and diversity in recommender systems: a survey". In: *ACM Transactions on Intelligent Systems and Technology* (2023).

[188] P. Ziegler. *Legacy: A Historical Account of an Early-Twentieth-Century International Scholarship Programme*. Yale University Press, 2008.

[189] John Zimmerman and Jodi Forlizzi. "Research Through Design in HCI". en. In: *Ways of Knowing in HCI*. Ed. by Judith S. Olson and Wendy A. Kellogg. New York, NY: Springer, 2014, pp. 167–189. URL: https://doi.org/10.1007/978-1-4939-0378-8_8 (visited on 11/03/2023).

[190] John Zimmerman and Jodi Forlizzi. "Speed Dating: Providing a Menu of Possible Futures". In: *She Ji: The Journal of Design, Economics, and Innovation* 3.1 (Mar. 2017), pp. 30–50.