Title

Neil Natarajan

New College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2024

Abstract

Artificially Intelligent (AI) systems are being implemented across a variety of fields, including the particularly sensitive field of Talent Identification (TI). With these implementations, existing ethical challenges are exacerbated and new challenges are posed. Many of these challenges centre on Diversity, Equity, and Inclusion (DEI). Here we contend that careful implementation of Interpretable Artificial Intelligence (IAI) and Decision Support Tools (DSTs) can alleviate existing DEI challenges, address new challenges, and improve TI professionals' decision-making.

We first analyse the existing state of AI, IAI, and DSTs, and their applications to TI problems. We define terms related to IAIDSTs and examine preconceptions about what makes specific IAIDSTs well-suited to this purposes. We then look at open DEI challenges in TI, including but not limited to those introduced or exacerbated by the use of AI in TI. Finally, we make a note of other systems designed to solve similar problems in TI.

We move on to examine a series of case studies wherein different types of IAIDSTs are used to address different DEI challenges in TI. In each case, we develop or examine the IAIDST used, note the use cases for which it is sufficient and the potential pitfalls in its implementation. We then demonstrate for a variety of types of IAIDSTs ways in which these systems can solve DEI issues in TI while addressing new complications and improving workflows and decision-making. In this examination, we look both at post-hoc explainable systems and intrinsically interpretable models.

Finally, we draw insights from these case studies and present prescriptions for how and when one might use the IAIDSTs discussed. We extrapolate these insights into more general recommendations for designing and implementing IAIDSTs in TI.

Title

Neil Natarajan

New College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2024

Acknowledgements

I would like to thank my advisors Reuben Binns and Nigel Shadbolt: Reuben, for his constant dedication to his students and willingness to run with any idea I came up with; Nigel, for his clear framing and pointed critiques. I would also like to thank my colleagues in Oxford Computer Science's HCAI theme, with special thanks to Sruthi, Ulrik, Jun, and Thomas. I would like to thank my family for their support, emotionally and intellectually, and my friends for my sanity.

Abstract

Artificially Intelligent (AI) systems are being implemented across a variety of fields, including the particularly sensitive field of Talent Identification (TI). With these implementations, existing ethical challenges are exacerbated and new challenges are posed. Many of these challenges centre on Diversity, Equity, and Inclusion (DEI). Here we contend that careful implementation of Interpretable Artificial Intelligence (IAI) and Decision Support Tools (DSTs) can alleviate existing DEI challenges, address new challenges, and improve TI professionals' decision-making.

We first analyse the existing state of AI, IAI, and DSTs, and their applications to TI problems. We define terms related to IAIDSTs and examine preconceptions about what makes specific IAIDSTs well-suited to this purposes. We then look at open DEI challenges in TI, including but not limited to those introduced or exacerbated by the use of AI in TI. Finally, we make a note of other systems designed to solve similar problems in TI.

We move on to examine a series of case studies wherein different types of IAIDSTs are used to address different DEI challenges in TI. In each case, we develop or examine the IAIDST used, note the use cases for which it is sufficient and the potential pitfalls in its implementation. We then demonstrate for a variety of types of IAIDSTs ways in which these systems can solve DEI issues in TI while addressing new complications and improving workflows and decision-making. In this examination, we look both at post-hoc explainable systems and intrinsically interpretable models.

Finally, we draw insights from these case studies and present prescriptions for how and when one might use the IAIDSTs discussed. We extrapolate these insights into more general recommendations for designing and implementing IAIDSTs in TI.

Contents

List of Abbreviations

AI . . . . . Artificial Intelligence

xAI . . . . . Explainable Artificial Intelligence

DEI . . . . . Diversity, Equity, and Inclusion.

IAI . . . . . Interpretable Artificial Intelligence

DST . . . . . Decision Support Tool

HitL . . . . . Human in the Loop

IAIDST . . . . Interpretable, Artificially Intelligent Decision Support Tool

TI . . . . . Talent Identification

1 Introduction

Contents

1.1 To-do

To-do

2 Methods

Contents

2.1 To-do

To-do

3

Limitations of Post-Hoc Explainable AI

Contents

3.1 Introduction

Increasingly, decisions affecting the lives of lay people are made by AI algorithms. And while these algorithms may be useful, they can also be dangerous. Users of AI systems desire an understanding of how these systems functions and why they yield the outputs they do, so that they may respond appropriately to the outputs [1]. When AI systems make decisions impacting the user, user insight into the decision-making process allows for recourse [2]. However, when decision-making systems involve both an AI and human component, user insight into AI outputs is crucial in both (1) inducing appropriate reliance or scepticism in the AI

as warranted and (2) optimising the overall performance of the decision-making system. This is especially true in high-stakes domains such as healthcare, finance, and criminal justice, where the decisions made by AI systems can have significant consequences for individuals [1–3].

The growing field of Explainable Artificial Intelligence (xAI) aims to develop methods for explaining algorithms. While some models are intrinsically interpretable [4], many more complicated machine learning algorithms are difficult to interpret. In order to explain these more complex models, various dedicated explanation algorithms have been proposed. Within the field of xAI, post-hoc, model-agnostic explainability methods remain popular due to their flexibility and ease-of-application [5]. Indeed, the ability of methods like SHAP, LIME, and Scoped Anchors to treat any model as a black-box allows for applications of explainability to models that would be otherwise inscrutable [6–8]. However, there are several known limitations that can impact their utility as IAIDSTs; in some cases, these limitations undermine the purpose of explanation, or even to deceive and mislead users [9].

In this Chapter, we explore a disconnect between user trust in an AI system and that system's trustworthiness as a result of post-hoc explainable AI tools. We begin by citing a series of papers yielding this result in a variety of contexts, and proceed to present our own work confirming that this disconnect persists in talent identification tasks. We then identify theories of when and why this disconnect occurs, and propose a series of experiments to test these theories.

In Chapter 4, we implement these experiments and present the results.

3.2 Trust and Trustworthiness in Explainable AI

User trust is central to the utility and limitations of any user-facing system. In the context of IAIDSTs, user trust in the system facilitates the system's adoption and use. Explainability is crucial here, as an oft-cited goal of transparency is increasing user trust in AI systems [9]. However, though trust is much-discussed, trustworthiness is often overlooked.

Jacovi et al. [10] set out rules for trust in an algorithmic context. They set out a basic philosophical analysis of trust (in the general, non-algorithmic sense): A trusts B if and only if A believes that B will act in A's best interest and accepts vulnerability to B's actions. Trust often has limited scope; typically, A will trust B regarding some particular actions or motivations, but not others.

In the context of trust in AI, this definition is only slightly changed. Jacovi et al. [10] characterise trust in an AI system by two properties: "the vulnerability of the user, and the ability to anticipate the impact of the AI model's decisions"; Vereschak, Bailly, and Caramiaux [11] similarly isolate three elements: "trust is linked to a situation of vulnerability and positive expectations, and is an attitude"; Lee and See [12] give a similar definition of trust: "An attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability". In all definitions, we see *vulnerability* emerge as a key concept, and we variably also see that trust is characterised by *uncertainty* and *expectations*.

An important distinction here can be drawn between whether someone or something is trusted and whether that trust is well-placed; i.e. it is worthy of trust [13]. In the machine context, Jacovi et al. argue, an algorithm is worthy of trust if and only if there exists some contract that the algorithm promises (or that the algorithm's creators and implementors promise) to uphold. They term this a 'contractual' model of trust. We adopt the nomenclature of contractual trust and use it to frame breakdowns in trust in specific post-hoc explanations.

3.3 A Series of Concerning Results

3.3.1 The Dangers of Transparency

It is often assumed in the xAI literature that transparency of AI systems is always desirable [5, 14]. However, Wang [9] dispel this claim. In particular, they demonstrate how, rather than empowering users and stakeholders, some implementations of algorithmic transparency serve the system's deployers, rather than its users, either through a false sense of understanding or the enforcement of norms and power structures. They examine the USA's FICO credit score, which

releases the factors that go into the score, the data sources for these factors, and general guidelines on how these factors are used. They find specific information on the use of these factors lacking, and conclude that, rather than empowering users to debate the ethics of certain decisions, this transparency only serves to enforce certain behaviours among the users [9].

Ustun, Spangher, and Liu [2] note a similar danger in transparency. They argue that explanations of AI systems making decisions about at end-users should empower those end users to alter the system's determinations. In particular, they introduce the concept of 'Actionable Recourse', which consists of a specific set of actions an end-user can take to change the AI's determination, and demonstrate how to calculate these for linear models. Key to actionable recourse is that the actions an explainee should take in response to the explanation are clear in the explanation [2].

We extend this concept to the context of IAIDSTs. Explanations aimed at decision-makers need not provide them *recourse*, as decision-makers already posses the ability to override the AI system. However, they should still provide *actionable* information, so that the explanations should contain information relevant to the decision-makers choice of what to do with the machine recommendation.

3.3.2 Black Box Explanations Increase User Trust in AI Systems

A number of studies exploring popular post-hoc, black-box explanation algorithms have found that these methods tend to increase overall user trust in the system being explained [9, 15, 16].

Ford, Kenny, and Keane [15] explore this effect in the context of a machine decision-maker with a human evaluator. They run a study examining the impact of post-hoc explanations-by-example and error-rates on people's perceptions of a black-box classifier classifying images from the MNIST dataset. They show that presenting 'case-based explanations' (a series of three important data points in the training of the model; elsewhere, we term this an 'influential instances'

explanation and use 'case-based' more broadly) lead participants to perceive misclassifications as more correct [15].

Jacobs et al. [16] extend this result to a human-in-the-loop context. They study the effect of machine recommendation on a clinician's ability to select antidepressant treatments, and find that providing clinicians an incorrect algorithmically generated treatment list lowers the accuracy of clinicians' own treatment lists. Notably, they do not isolate this result to the presence of an explanation, but rather demonstrate how two explainable AI systems harm clinician antidepressant selection relative to a placebo group. However, they also find that feature-based explanations mislead clinicians more than heuristic-based explanations. Modern best practices outlined by Miller [14] prefer the usage of feature-based explanations, so we find this particularly alarming [16].

McCradden [17] responds to Jacobs et al. [16] questioning the role of accuracy in explanations of IAIDSTs. They note the focus of Jacobs et al. [16] on improving *accuracy* of the treatment lists, and aim to do this with IAIDSTs. However, though these systems may optimise for accuracy, they notably fail (at least in these specific instances) to facilitate clinicians' ability to *help patients*, which is a clinicians' primary goal [17].

In a broader context, while 'helping patients' doesn't describe the goal of all IAIDSTs, neither does 'improving accuracy'. And, though increasing decision-maker trust regardless of system veracity may improve overall accuracy, it may hamper the broader goal of the decision-making system.

3.3.3 Impacts on Talent Identification

It should be noted that these effects are not universal. Mohseni et al. [18] measure user trust in an AI fake news detection system over time, and find the profile of the users to be just as important as the type of explanation. Though they demonstrate that different explanatory conditions have differential effects on which profile participants are likely to exhibit, they also cluster trust into five profiles by user: consistent over-trust, consistent under-trust, consistent trust, trust gains

continually, and trust decreases continually. These profiles suggest that the same explanations will impact different groups performing different tasks differently. Thus, while an IAIDST may lead clinicians selecting antidepressants astray, it may have no effect (or even the opposite effect) on talent identification tasks [18].

In the next section, we undertake a series of experiments to test the implications of these limitations for talent identification.

3.4 Misleading Explanations of AI Outputs in Talent Identification

To-do: *Explainable AI (xAI) methods are often motivated by the need to increase user trust in AI systems. However, increased trust may not always be desirable. While too little trust might lead people to neglect AI systems when they are correct, too much trust can also be a risk. Explanations which encourage misplaced trust in incorrect AI outputs might mislead humans into making bad decisions. We investigate the extent to which three different xAI methods create too much trust in their underlying AI systems. Examining Shapley-based Additive Explanations (SHAP), Scoped Anchors, and a confidence-based explanation that presents the model's confidence in an output, we conduct experiments where participants perform tasks with xAI assistance. We find that when the system is wrong, SHAP or Confidence explanations still increase trust. Scoped Anchors explanations, by contrast, increase participants' confidence in their own estimations regardless of the system's correctness. We discuss implications for the design and deployment of xAI in HitL tasks.*

3.5 Trust Explanations to Do What They Say

To-do: *Trusting an algorithm without cause may lead to abuse, and mistrusting it may similarly lead to disuse. Trust in an AI is only desirable if it is warranted; thus, calibrating trust is critical to ensuring appropriate use. In the name of calibrating trust appropriately, AI developers should provide contracts specifying use cases in which an algorithm can and cannot be trusted. Automated explanation of AI*

outputs is often touted as a method by which trust can be built in the algorithm. However, automated explanations arise from algorithms themselves, so trust in these explanations is similarly only desirable if it is warranted. Developers of algorithms explaining AI outputs (xAI algorithms) should provide similar contracts, which should specify use cases in which an explanation can and cannot be trusted.

4

Using Post-Hoc Explainable AI to Identify Talent

Contents

4.1 Introduction

To-do

4.2 Explaining Applicant Resumes with Highlighting and Blinding

We develop a white-box explanation algorithm for highlighting the resumes of job applicants and scoring them. We use only the highlights themselves, opting for an evaluative AI approach, and provide a contract for the intended use case of our

algorithm. We then implement this tool in the hiring process of an organisation and conduct a field study on its impact on hiring decision-making. We discover...

(To-do)

4.3 Explaining an Applicant Scoring Algorithm with Shapley Values

We develop a SHAP-based procedure for explaining the decisions made by an applicant scoring algorithm used in a talent investment organisation. We take into account the original functions of SHAP explanations and provide a contract for the intended use case of our procedure: to systematically reveal insights about the underlying algorithm itself, help us better understand counterintuitive results, and instruct us on where this algorithm should be modified. We explore decisions made by the talent investment organisation and analyse the applicant scoring algorithm using the SHAP tool. We discover... Finally, we discuss the potential for extending this tool to use in decision-making and note strategies to combat mis-calibrated trust.

(To-do)

4.4 Detecting Generative AI Usage in Application Essays

Student use of generative AI in essay-writing creates new challenges for education in the marking of essays and essay-based selection for scholarships, fellowships, and universities. In theory, software purporting to detect AI-generated content can act as a post-hoc explanation of generated content, and thus offers a plausible solution to educators inquiring about the use of generative AI in student essays. In practice, little research has attempted to validate such solutions in real-world situations or to examine their limitations. We present an empirical case study exploring the efficacy and implications of using one such detection product, GPTZero, in the selection process for a program looking for talented young people from around the world. We observe that GPTZero does not perform sufficiently well for the program to

disqualify applicants on its basis alone. Also, GPTZero's scores are heterogeneously biased across geographical and gender groups. However, we find GPTZero accurate enough to conduct useful aggregate analyses of potential vulnerability to AI-enabled attacks on the integrity of the program's application process. We find evidence for only limited use of AI-generated text in the program's latest application cycle and no evidence partner organizations used AI-generated text at scale to defraud the program's referral program.

(To-do)

5

Using Intrinsically Interpretable Models to Identify Talent

Contents

5.1 Introduction

To-do

5.2 What Do We Mean When We Talk About 'Diversity'?

Talent Identification professionals frequently speak of diversity as a selection desideratum. This is variously used in the contexts of selected cohorts and selected individuals. Several definitions of diversity exist in the literature, but none encapsulate what TI professionals mean when they talk about diversity. We

conduct a series of interviews with TI professionals investigating what they mean by "diversity", how they measure diversity, and why diversity is important. We conduct a thematic analysis of these interviews and report several themes related to the meaning of diversity. Finally, we discuss the significance of these insights and their applications to the development of purpose-built decision assistants designed to help TI professionals make diversity-conscious selection decisions.

(To-do)

5.3 A Possibility Frontier Approach to Diverse Talent Selection

Organisations (e.g., talent investment programs, schools, firms) are perennially interested in selecting cohorts of talented people. And organizations are increasingly interested in selecting diverse cohorts. Except in trivial cases, measuring the tradeoff between cohort diversity and talent is computationally difficult. Thus, organizations are presently unable to make Pareto-efficient decisions about these tradeoffs. We build on disparate understandings of diversity and introduce an algorithm that approximates the upper bound on cohort talent and diversity, as measured by one of a variety of target functions capturing different desiderata. We call this object the selection possibility frontier (SPF). We then use the SPF to assess the efficiency of the selection of a talent investment program run in 2021, 2022, and 2023. We show that, in the 2021 and 2022 cycles, the program selected cohorts of finalists that could have been better along both diversity and talent dimensions (i.e., considering only these dimensions as we subsequently calculated them, they are Pareto-inferior cohorts). But, when given access to our approximation of the SPF in the 2023 cycle, the program adjusted decisions and selected a cohort on the SPF.

(To-do)

5.4 Co-designing Interpretable, Artificially Intelligent Decision Support Tools for Understanding Diversity

The SPF represents one approach to resolving some of the concerns TI professionals have when discussing diversity. We interview TI professionals to understand what they would desire from tools designed to help them better address diversity concerns in selection. Furthermore, in addition to adapting the SPF methodology into a prototype designed to address specific TI professional concerns, we develop two other prototypes based on these interviews. Then, we run a scenario speed-dating exercise in order to understand the strengths and weaknesses of each approach. We evaluate each approach by how well it solves the problem it is intended to solve. Finally, we adapt these insights into a series of guidelines for how to design tools

(To-do)

6 Discussion

Contents

6.1 Implications and Recommendations

Here we analyse the implications of using IAIDSTs in talent identification. We first look at the strengths and limitations of post-hoc explanations and prescribe appropriate use cases for these types of IAIDSTs. Second, we examine the strengths and weaknesses of intrinsically interpretable models, including purpose-built models, and make similar prescriptions. Finally, we distil our insights in purpose-building IAIDSTs for TI professionals and extrapolate to general recommendations.

6.2 Limitations and Future Work

We note the limitations of this research method. We examine first limitations of individual case studies and note how other methodologies might address these limitations. We then raise concerns about our research's external validity, especially

with regard to our recommendations and prescriptions. We address these concerns and note ways in which this research should not be used. Finally, we highlight areas in which researchers may continue this work or expand upon it.

6.3 Conclusion

To-do

Appendices

References

[1] Reuben Binns. "Human Judgment in algorithmic loops: Individual justice and automated decision-making". en. In: *Regulation & Governance* 16.1 (2022). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/rego.12358, pp. 197–211. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/rego.12358 (visited on 04/14/2022).

[2] 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).

[3] 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.

[4] Cynthia Rudin et al. "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges". In: *arXiv:2103.11251 [cs, stat]* (July 2021). arXiv: 2103.11251. URL: http://arxiv.org/abs/2103.11251 (visited on 04/14/2022).

[5] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2019.

[6] 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).

[7] 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).

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Anchors: High Precision Model-Agnostic Explanations". en. In: *AAAI*. 2018, p. 9.

[9] Hao Wang. "Transparency as Manipulation? Uncovering the Disciplinary Power of Algorithmic Transparency". en. In: *Philosophy & Technology* 35.3 (Sept. 2022), p. 69. URL: https://link.springer.com/10.1007/s13347-022-00564-w (visited on 08/04/2022).

[10] Alon Jacovi et al. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. New York, NY, USA: Association for Computing Machinery, Mar. 2021, pp. 624–635. URL: https://doi.org/10.1145/3442188.3445923 (visited on 10/15/2021).

[11] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. "How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies". In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (Oct. 2021), 327:1–327:39. URL: https://doi.org/10.1145/3476068 (visited on 04/14/2022).

[12] John D. Lee and Katrina A. See. "Trust in Automation: Designing for Appropriate Reliance". In: *Human Factors* 46.1 (Mar. 2004). Publisher: SAGE Publications Inc, pp. 50–80. URL: https://journals.sagepub.com/doi/abs/10.1518/hfes.46.1.50_30392 (visited on 09/16/2022).

[13] Russell Hardin. *Trust and trustworthiness*. Trust and trustworthiness. New York, NY, US: Russell Sage Foundation, 2002, pp. xxi, 234.

[14] Tim Miller. "Explanation in Artificial Intelligence: Insights from the Social Sciences". In: *CoRR* abs/1706.07269 (2017). _eprint: 1706.07269. URL: http://arxiv.org/abs/1706.07269.

[15] 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).

[16] 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).

[17] Melissa D. McCradden. "When is accuracy off-target?" en. In: *Translational Psychiatry* 11.1 (June 2021), p. 369. URL: http://www.nature.com/articles/s41398-021-01479-4 (visited on 09/16/2021).

[18] Sina Mohseni et al. "Trust Evolution Over Time in Explainable AI for Fake News Detection". en. In: p. 4.