# 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 |

See Appendix **??** for definitions of terms used in this Thesis, including these abbreviations.

# 1
# Introduction

## Contents

Since early theorising about Artificial Intelligence (AI), experts have noted the transformative potential of such technology. Now, with the popular advent of Machine Learning (ML) models and transformer architecture in particular, we see rapid mass-market adoption of a variety of AI tools. In many fields which deal with sensitive decision-making, this adoption introduces new ethical challenges. In the fields of talent and Talent Identification (TI) in particular, many of these ethical challenges centre on Diversity, Equity, and Inclusion (DEI) or related notions such as equality or justice. In this section, we lay out the case for the application of AI systems (specifically interpretable AI decision assistants) to DEI issues in TI, paying special attention to the potential pitfalls and broader implications of our proposal.

## 1.1 Motivation

This is where I explain why I'm doing this research. I explain why care about talent identification and how IAIDSTs can solve problems here. I respond to the notion that we should simply not use AI here, and highlight problems in the field already. I explain why this is a problem for Computer Science as well as for Talent Science (tl;dr: additional reasons to care about things xai overlooks because of the application domain)

## 1.2 Positions

### 1.2.1 Trust Explanations to Do What They Say

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. Both unwarranted trust in such an algorithm may lead to wrong and damaging decisions, while unwarranted mistrust in such an algorithm may lead to disuse. Neither outcome is favourable. Hardin draws an important distinction between whether someone or something is trusted and whether that trust is well-placed; i.e. it is worthy of trust [**hardin_trust_2002**].

It is clear, then, that trust in AI algorithms should be *calibrated*, so that users are led to trust trustworthy AI systems and distrust untrustworthy AI systems. **jacovi_formalizing_2021** propose that trust in AI systems can be understood in terms of a contract between the system and the trustor.

**jacovi_formalizing_2021** define a model of human-AI trust resting on two key properties: the *vulnerability* of the user to the model and the user's ability to *anticipate* the impact of the AI model's decisions. In the human context, person $A$ trusts person $B$ if and only if $A$ believes that $B$ will act in $A$'s best interest, and $A$ accepts vulnerability to $B$'s actions. In the machine context, we do not always expect the machine to act in our best interests. Instead, user $U$ trusts AI model $M$ if and only if $U$ can anticipate and accepts vulnerability to $M$'s actions [**jacovi_formalizing_2021**].

Moreover, trust often does not have a blanket scope; typically, $U$ will trust $M$ regarding some particular actions or range of actions, though a broader trust will include many such actions. In the algorithmic context, this scope is clearly limited – unlike humans, trust in algorithms should never be broad; warranted trust is always scoped to a region in which the algorithm's actions can be anticipated, and in which users might reasonably accept vulnerability to these actions. Generally, this scope is limited to some subsection of the intended use cases of the AI system. Trust placed in an AI system to do something it was not intended to do is often unwarranted; trust placed in an AI system to do something it does not claim to do is always unwarranted. Thus, for an algorithm to be trustworthy in a given scope, that algorithm should demonstrate both that a user can anticipate behaviour in that scope and that the anticipated behaviour is such that users might accept vulnerability to the algorithm. We call this demonstration a *contract*, and call this sort of trust *contractual trust* [**jacovi_formalizing_2021**].

Following this framing, the extent to which an algorithm warrants trust is modulated by the extent to which it adheres to its contract. Therefore, when the developers of an algorithm provide a contract regarding the intended use of an algorithm, we can evaluate the trustworthiness of an algorithm by evaluating adherence to the contract.

One method of evaluating adherence to contract comes from a user observing the AI algorithm's reasoning process by way of an explanation or an interpretation. However, unlike human decision-makers, few algorithms are inherently capable of explaining their reasoning. The growing field of Explainable Artificial Intelligence (xAI) aims to develop methods for explaining the reasoning algorithms, often with a broad goal of increasing warranted trust in algorithms. However, though it is clear that these algorithms often increase trust in algorithms, it is not always clear that this trust is warranted, as demonstrated by **jacobs_how_2021**. Thus, it seems, there are times where even an explanation of an AI algorithm should be distrusted.

### Trust in Explanations of AI Algorithms

Explanation algorithms help us determine whether to trust AI algorithms, but only if we trust the explanation methods. But when can we trust an explanation algorithm? And what are we trusting it to do? The answer that we are trusting these algorithms to *explain* AI systems is insufficient, because what it means to explain is unspecified. AI explainers can be put to a number of different uses, and different algorithms should be trusted for different uses; contracting to behave appropriately in all of these uses is infeasible (an end-user demands a different explanation than a domain expert), so explanations methods should contract to provide only a particular type of explanation.

Much like AI models themselves, we contend that xAI algorithms should be trusted to uphold specific contracts with respect to the ways in which they are used. For example, a model like *recourse*, developed in **ustun_actionable_2019**, designed to informs end-users of what must be done to change their determination, should not be trusted to report errors in model or to point out the most important features. Similarly, a model like *Scoped Anchors*, developed in **ribeiro_anchors_2018**, designed to simplify predictions into rule-based approximations, should not be trusted to provide recourse information.

We also contend that xAI methods should be evaluated on whether they can be trusted to do what they say. That is, a good xAI method is one that fulfills its intended use case. Much like trust in AI algorithms, trust in explanations of AI algorithms is contractual; xAI methods should be evaluated in terms of the extent to which they uphold the terms of a contract between the explainer and explainee; and an explainee's trust should be calibrated accordingly. We should not trust an explanation algorithm to do something it has not promised to do.

The absence of contracts is not a mere conceptual problem; it creates a problematic dialectic and hinders effective critique of xAI methods. To demonstrate this, we consider two particular kinds of AI explanations: SHAP explanations, introduced in **lundberg_unified_2017**, and counterfactual explanations, introduced in **wachter_counterfactual_2017**. Both papers focus on the mathematical properties of the explanation algorithm introduce, but neither makes clear what they contend a good explanation consists in or specifies a circumscribed set of use cases for their methods. We consider two evaluation articles: **kumar_problems_2020**'s evaluation of the SHAP method, and **barocas_hidden_2020**'s evaluation of counterfactual explanations. Both articles rely on similar notions regarding the purpose of explanations – frameworks that the authors of SHAP and counterfactual explanations do not make clear that they subscribe to. For instance, **barocas_hidden_2020**'s critique counterfactual explanations on the grounds that they are not useful in providing users with actionable information. Similarly, **kumar_problems_2020** argue that SHAP cannot be used to inform users' actions.

We contend that, like trust in AI algorithms, trust in AI explanation algorithms is composed of an ability to anticipate the algorithm and an acceptance of vulnerability to the algorithm's actions. In both cases, this trust is only desirable if it is warranted. The scope of the trust, in both cases, should be clearly enumerated in a contract, and AI and explanation algorithms alike should be evaluated for trustworthiness within this scope.

## 1.3 Research Questions [WIP]

I enumerate the research questions here, then explain how I intend to answer them.

## 1.4 Research Outputs [WIP]

I enumerate the answers to these research questions in the form of research outputs.

## 1.5 Thesis Structure [WIP]

I explain the flow of everything and why I've separated the chapters the way I have. E.g., "in chapter 3, we answer X. The reason this ties to the central themes of the research is Y. This leads us into chapter 4, where we answer Z"

## 1.6 Publications [WIP]

I list the publications that have come out of this research.

# 2 Background

In this chapter we examine: new advancements in IAI, open problems in TI and their existing solutions, and human-centric methodologies that might help us applying AI to TI. We first engage in a survey of the field of IAI, including the types of articles prevalent in the field, the prominent paradigms for interpretability and explanation, and several key concepts in the development of new algorithms. Next, we examine open problems in diverse, equitable, and inclusive TI and note pre-existing solutions from prior works. Finally, we examine human-centred methodologies that we may draw on in crafting our own solutions. We further note that some problems are not well-suited to resolution through explanation, and suggest a variety of other solutions that could be attempted in address of these problems.

## Contents

## 2.1 Interpretable Artificial Intelligence

### 2.1.1 Types of IAI Publications

The field of IAI can, broadly, be divided into four main bodies of research: reviews, methods, notions, and evaluations [**vilone_explainable_2020**]. Review articles are either systematic investigations or literature reviews. Method articles introduce new explanation methods. Notion articles focus on defining notions or concepts related to explainability. Evaluation articles evaluate existing notions or paradigms.

#### Review Articles

Review articles offer meta-commentary on other articles listed below. A review article is so-characterised because its primary contribution is one of aggregation. Thus, a review article might also be a method, notion, or evaluation article, so long as it primarily draws from these articles. This paper is, primarily, a review article.

#### Method Articles

In most method articles, the task of explanation is taken to be fairly well-defined, and the research questions deal in how to most effectively generate that explanation.

Consider two particular method articles: SHAP explanations, introduced by **lundberg_unified_2017**, and counterfactual explanations, introduced by **wachter_counterfactual_**
In both papers, we see heavy treatment of the mathematical properties of the explanation algorithm introduced. However, we see relatively light treatment of

what the authors take to be a good explanation or proposed use cases for the novel methods. These concepts are instead the domain of notion articles.

### Notion Articles

Where method articles deal with individual methods, notion articles deal with theory underlying a variety of methods. For example, **miller_explanation_2017** discussion of the importance of principles of explanation from the social sciences is a notion article.

### Evaluation Articles

Evaluation articles consider which notions and methods are best-suited to produce the best explanations. Evaluations of method articles, in particular, often compare them to notions introduced elsewhere, and where critiques are found, they are often critiques that the method fails to meet a standard set by the notion.

For the two method articles discussed above, we present two evaluation articles: **kumar_problems_2020** evaluation of the SHAP method, and **barocas_hidden_2020** evaluation of counterfactual explanations. Both articles rely on similar notions regarding the purpose of explanations, and critique the methods for failing to adhere to a notion. For instance, **barocas_hidden_2020** assume that counterfactual explanations are given to provide users actionable information. Similarly, **kumar_problems_2020** argue that SHAP cannot be used to inform users' actions.

## 2.1.2 A Taxonomy of IAI Methods

AI model explanations can broadly be categorised into two subgroups: intrinsically interpretable models and post-hoc explanations [**molnar_interpretable_2019**]. Though intrinsically interpretable models offer explanations and are meaningfully classed as forms of "interpretable AI", they do not make use of the major developments in "explainable AI" [**molnar_interpretable_2019**], since they are already interpretable.

Post-hoc methods can be further separated into global methods, which explain entire models, and local methods, which explain individual decisions [**molnar_interpretable_2019**]. When assessing explanation methods, one can distinguish between local tests and global tests of model explanations [**molnar_interpretable_2019**]. Local tests concentrate on the impact of the explanations at the level of a single prediction, global tests focus on the impact of explanations on overall model trust.

Rather than delineating between intrinsic and post-hoc or global and local, we could instead delineate by output type [**friedrich_taxonomy_2011**]. Some interpretability devices offer feature importance statistics. Others offer explanations by contrast to another datapoint. Yet others offer rules that guide decision-making in regard to an individual case case.

### 2.1.3 Use Case

Consider the following potential use cases for xAI methods: to evaluate whether the AI is operating as intended, to examine an AI for bias and unfairness, to provide recourse, to provide a rule that would guarantee equality. This list is not exhaustive, and yet, even for these three use cases, explanations that satisfy some will not satisfy others [**natarajan_trust_2023**].

### 2.1.4 Theories of Explanation

In deciding which type of explanation will best suit a particular use case, we must consider the theory of explanation that underlies the explanation, and the desiderata that that theory demands of its explanations.

**miller_explanation_2017** summarises key findings from Social Sciences related to explanation. **miller_explanation_2017** works from a causal theory of explanation; he outlines and examines the theory, noting challenges to it, but does not consider a rejection of the theory. According to the theory, explanations are sought in response to counterfactual cases, i.e. the question asked is not of the form "Why $P$" but rather "Why $P$ instead of $Q$" [**miller_explanation_2017**]. Such counterfactual theories of causality all argue that cause is a matter of what would have happened if the cause had not happened. Though they disagree on precisely how to model these counterfactual cases, many leading philosophical theories of causality agree that causality is best understood in terms of counterfactuals. Hume modeled these counterfactuals in terms of causes and events: in actuality, cause $C$ caused event $E$ to occur; but if cause $C'$ had replaced cause $C$, $E$ would not have occurred. We can call this sort of counterfactual a cause-counterfactual [**miller_explanation_2017**]. Under this conception, explanations should present cause-counterfactuals as clearly as possible.

**miller_explanation_2017** also argues that explanations should be contrastive and selective. Contrastive explanations make use of counterfactuals, but these counterfactuals are not cause-counterfactuals. Instead, they are event-counterfactuals. In other words, contrastive explanations consider alternative events $E'$, and explain the data $E$ in relation to $E'$. Selective explanations, on the other hand, are ones that do not consist of a complete cause, but rather select one or two causes from the complete cause.

**miller_explanation_2017** presents a powerful application of a theory of explanation to the field of xAI. However, causal theories of explanation are far from the only type of explanatory theories. Indeed, as we noted above with numerical evidence, there are theories of explanation who's desiderata contradict those of **miller_explanation_2017**'s causal theory [**woodward_scientific_2021**].

In what follows, I present **woodward_scientific_2021**'s theory, the statistical relevance model of explanation. The statistical relevance theory rests heavily on the notion of statistical relevance. For events $A$, $B$, and $C$, we say $C$ is statistically relevant to $B$ in $A$ if and only if $P(B|A \wedge C) \neq P(B|A)$. I.e., an explanation is a member $C_i$ of a homogeneous partition $C$ of properties: that is, a set of properties that are exclusive and exhaustive of $A$, where there are no statistically relevant properties $D$ to $B$ in $A \wedge C_i$. The explanation also consists of $P(B|A)$, the probability $P(B|A \wedge C_i)$ of each cell withing the partition, and which of the $C_i$ contains the desired point to be explained, $x$ [**woodward_scientific_2021**].

On both accounts , explanations have an explainer and an explainee. Explanations in everyday use are an interaction. Unlike most modern algorithmic explanation methods, an explanation given by a human is often given in the form of a conversation. Thus, it obeys the rules of communication using language. In other words, these explanations are social. We may therefore apply theories governing social interactions, such as **Grice_1975**'s conversational maxims, to explanations [**miller_explanation_2017**].

**Grice_1975**'s four categories of conversational maxims are: quality, quantity, relation, and manner. That is, the quality of information conveyed in a cooperative conversation should be high (information should be likely and justifiable). The quantity of information should be neither too little nor too much. The information should be related to the conversation. The information should be conveyed in an appropriate manner. Note that the quantity bounds overlap quite heavily with selectivity.

### 2.1.5 Trust

Another important notion related to xAI is trust. Indeed, much of the xAI literature suggests that the purpose of xAI is to increase trust in AI. This is apparent from the titles of canonical papers in the field, such as **ribeiro_why_2016** and **pieters_explanation_2011**. So what does it mean to trust an AI system? A basic philosophical analysis of what trust consists in (in the general, non-algorithmic sense) is as follows: $A$ trusts $B$ if and only if $A$ believes that $B$ will act in $A$'s best interest, and $A$ accepts vulnerability to $B$'s actions [**jacovi_formalizing_2021**]. Moreover, trust often does not have a blanket scope; typically, $A$ will trust $B$ regarding some particular actions or motivations.

An important distinction here can be drawn between whether someone or something is trusted and whether that trust is warranted; i.e. it is worthy of trust [**hardin_trust_2002**]. In the context of AI, according to **jacovi_formalizing_2021**, we should only ever trust AI systems to fulfil their contracts, but some AI systems are not worthy of even this trust. Ideally, trust in a system should relate directly to that system's trustworthiness – we should place trust in any system when it warrants trust, but should not place unwarranted trust in any system. The problem of appropriately calibrating trust, however, is yet unsolved.

There are many methods that aim to calibrate trust in an AI algorithm. We could give some access to the patterns that distinguish correct and incorrect cases. The better some end-user is able to distinguish these, the better they know when they can trust the model to be correct. Knowledge of the performance of a model grants access to the patterns distinguishing correct and incorrect cases. Facts about the bias of the model grant access as well, as does information about the performance of the model on a subset of data. Explanations can be seen as one way of imparting such information. Because of this, algorithm explanations allow explainees to determine when an algorithm is trustworthy.

### 2.1.6 Relevant Evaluations of XAI

There is a growing body of studies that empirically evaluate xAI methods. In some cases, an xAI method is evaluated not with human subject experiments but rather with some formally defined proxy for interpretability [**doshi-velez_towards_2017**]. However, most evaluations involve humans using an AI system to perform some task, and the effect of an xAI method on task performance is measured. For our purposes, namely the measurement of end-user trust and reliance, we need to measure lay people using a system in a simplified but realistic task.

It should also be noted that there is no standard evaluation or set of benchmarks explainability methods are judged by [**doshi-velez_towards_2017**]. Indeed, though there are proposed methods for evaluating attribution-based explanations and model-based explanations, there are none for evaluating example-based explanations [**markus_role_2021**]. As a result, there exist a variety of study designs, conditions, and variables measured by human-centred evaluation papers. We list some such studies here.

**ford_play_2020** run a study where they examine the impact of post-hoc explanations-by-example and error-rates on people's perceptions of a black-box classifier . They show that case-based explanations lead participants to perceive miss-classifications as more correct. In the case of their study, a case-based explanation is a series of three important data points in the training of the model (in other words, an "influential instances" explanation). They also show that classifiers with higher error rates lead participants to perceive the models as less trustworthy. They show participants a series of machine classifications of the MNIST dataset, where the model either correctly labels the data or commits an alternate labelling error. They then task the participants with rating correctness and reasonableness of each classification on a 5-point scale. Finally, at the end, the participants filled out global correctness and reasonableness, alongside global trust forms. The design was a 2x3 design, with explanations present or absent, and three accuracy levels. Participants report miss-classifications as being more correct when given a case-based explanation. They did not find significant findings for reasonableness, trust, or satisfaction across explanation-present or explanation-absent conditions. They use a MANOVA design to account for differences across conditions, but use a between-subjects design [**ford_play_2020**].

**jacobs_how_2021** run a study in the medical diagnosis context, where participants are given a vignette of a patient and, in the experiment condition, a recommended treatment list and explanation. Participants were asked to make an antidepressant treatment selection, to rate their confidence, and to indicate the utility of the model on their decision (it appears on a 5-point scale). The study had statistically significant results only with respect to accuracy (measured as an average of 0s and 1s), not confidence or perceived utility. In the accuracy case, this indicated that an algorithmically generated treatment list, when wrong, would mislead the participants and lower their accuracy. In the confidence and utility cases, this would indicate either that accuracy is affected, but not confidence or utility, or that the Likert scale measurement method used in Jacobs et al.'s study is a weaker measurement tool than the accuracy. Furthermore, these findings are not isolated to the explanation. Rather, Jacobs et al. consider the recommendation and the explanation together. For example, they find that feature-based explanations, paired with incorrect recommendations, lowered accuracy compared to the baseline condition (no recommendation) [**jacobs_how_2021**].

**bansal_does_2021** perform another similar study, looking at human-AI cooperation in a context where they have comparable performance. They use sentiment classification as their task. Furthermore, they measure effects with a variety of explanations. They find that explanations inspire increased trust in AI systems regardless of the correctness of the model. As explanation, they use a highlighting of the top predicted sentiment classes, alongside a highlighting of important words. Finally, they test expert-generated explanations to serve as an upper bound, as they found that humans were often confused when machine explanations made little sense. They note that explanations make the user more likely to have high accuracy when the AI is correct, but more likely to have low accuracy when the AI is incorrect. This would indicate that AI recommendations with explanations are capable of fostering mistrust, as they foster both overtrust and under trust [**bansal_does_2021**].

**mohseni_trust_nodate** measure user trust in an AI fake news detection system over time. They design a study in which users are shown true and fake news articles. The users are asked at three different points (each one additional third through the study) what their perceived accuracy of the explainable AI system is. They had three different conditions, segmented by type of explanation. There was a baseline condition with no explanation and two xAI conditions. They show that the user trust can be clustered into five profiles: consistent overtrust, consistent under trust, consistent trust, trust gains continually, and trust decreases continually. They show that more participants from the no explanation and attribute explanation conditions were consistently gaining trust, and that more participants from the attention explanation condition overshot their second perceived accuracy measurement [**mohseni_trust_nodate**].

## 2.2 Open Problems in Diverse, Equitable, and Inclusive Talent Identification

We highlight several open problems in TI that touch on concerns of DEI. To begin with, we define DEI alongside related concepts such as JEDI, EDI, DEIB, and IDEA. Then we spotlight the potential equity challenges posed by the integration of DSTs into TI workflows. In particular, we note that any bias and fairness issues built into a DST will create bias and fairness issues for the corresponding TI workflow. However, we note that problems of bias and fairness in TI exist even in the absence of DSTs. We move on to discuss fairness issues posed by applicant use of Generative AI. Finally, we note that TI professionals consistently struggle to balance selecting talented individuals with constructing a diverse cohort and note the implications for equity that this struggle brings.

### 2.2.1 Bias

#### The Problem of Bias

We can define the problem of bias as the liability of systems to exhibit different behaviour on groups segregated across a "special" partition. "Special" here refers to partitions across protected classes, or partitions across which we would expect a system to perform equally. For example, gender and race are usually special partitions in the context of talent identification, but academic ability and teachability are not.

To illustrate the issue of bias in AI systems, we turn to **mattu_how_nodate**'s analysis of the COMPAS system. The COMPAS system was an expert-system used to predict recidivism of criminals in the US. When defendants accused of crimes were booked in jail, they responded to a COMPAS questionnaire. These answers were then fed into the COMPAS software to produce predictions of the "Risk of Recidivism" and "Risk of Violent Recidivism." These predictions were used, among other things, to determine if a defendant could be set bail (as likelihood of reoffence is a significant factor in the setting of bail). When **mattu_how_nodate** analysed the COMPAS tool's predictions relative to actual recidivism rates, they found that the recidivism predictions had roughly equal accuracy for White and

Black defendants, but that, when errors were made, these errors acted in very different directions. That is, the Black defendants were more likely to be falsely predicted to reoffend, and the White defendants were more likely to be falsely predicted to not reoffend. As a result, the implementation of the COMPAS system had significantly greater negative externalities on the Black community than on the White community [**mattu_how_nodate**].

**barocas_big_2016** illustrate how data mining yields exactly this sort of bias. In the example of the COMPAS set, much of this bias comes from the underlying training data, but the software's predictions were, on average, *more* biased than the underlying data, as the process the software used to optimise its predictions exacerbated any biases in the underlying data [**barocas_big_2016**].

### How Modifying Training Data Addresses Bias

In some cases, algorithmic bias introduced via latent biases in the training data can be corrected rather simply by altering training data before the training process. One case of bias arises when one group across a special partition is much smaller than other groups [**barocas_big_2016**]. Due to this sample size imbalance, optimising for accuracy across the total group leads to overemphasis on accuracy on the large group, and allows for marginal improvements in performance on the larger groups at the cost of performance on the smaller groups. While we might address this problem with explainability methods, in this case, the most straightforward solution rather involves increasing the attention paid to the small groups in training, either by increasing the rate at which we sample from this group, or by increasing the weight applied to each sample [**barocas_big_2016**].

However, though simple, this solution is not always effective. Sometimes, for example, biases are introduced in the content of the training data, rather than their distribution. In this case, heterogenously biases errors in the variable to be predicted induce those same biases in the predictive algorithm. In this case, the solution is to correct the biases in the data itself, but this is often either prohibitively difficult, or outright impossible [**barocas_big_2016**].

### How IAI Addresses Bias

When bias cannot be removed, it instead might be managed. Where biased behavioural patterns exist in an AI system, simply making these systems more interpretable will not alter these patterns themselves. However, with humans in the loop, spotlighting algorithmic bias might allow humans to address it.

Consider the COMPAS example again, except suppose now that COMPAS implements a counterfactual explanation, which points to a similar hypothetical input that would have a different output. Suppose a defendant has a high calculated risk of reoffense. The actual defendant's race is "black", but the counterfactual explanation reveals that a similar white defendant would have had a low calculated risk of reoffense. That is, COMPAS is, in this case, clearly discriminating based on race. If this were used in a courtroom setting, a judge might see this information and subsequently choose to disregard the prediction yielded by COMPAS. In this way, an explanation could indicate that an algorithm was biased and encourage scrutiny of that algorithm [**mothilal_explaining_2019**, **wachter_counterfactual_2017**].

Furthermore, in certain special circumstances, implementing techniques from IAI at training time can actually remove bias. For example, we could imagine imposing explanation-based constraints on a model at the time of training. One such constraint would be **wang_deontological_2020**'s 'monotonicity' constraint , which would require that a particular input i relate monotonically to the model's output. Thus, if feature sets X and X' differ only in i such that i in X' is greater than i in X, and M is a model that is monotonic in i, then M(X') should be greater than M(X). This constraint would prevent a model from making decisions that penalise applicants based on, for example, membership in a minority group.

## 2.2.2 Fairness

### The Problem of Fairness

Fairness is closely tied to, though still distinct from, Bias. Where bias deals with groups, fairness deals with individuals. We can define the problem of fairness as the liability of AI systems to exhibit different behaviour on different individuals that are similar in relevant dimensions. We can note this follows the given principle, entitled the “similar treatment” principle: if two individuals are similar across relevant dimensions, they should be treated similarly [**Fleisher_2021**, **dwork_fairness_2012**].

Consider again the COMPAS example. Where the disproportional treatment of groups is a form of bias, an individual decision yielding a false positive in one individual, and a false negative in another individual is a problem of fairness. We can call this form of fairness “Individual Fairness”, and contrast it with other forms of fairness such as statistical parity fairness, where a system is fair if and only if it achieves parity between groups [**Fleisher_2021**, **dwork_fairness_2012**]. We focus on individual fairness as a particularly important definition of fairness, and note that statistical parity fairness is more closely bias, as we have defined it. However, note that individual fairness is not a privileged form of fairness. **Fleisher_2021** argues this, and notes that individual fairness is undesirable when it conflicts with other notions of fairness. Thus, we also note that there are more general notions of fairness that we might address as well.

### 2.2.3 Addressing Fairness

Much like bias, fairness is not address directly via interpretability alone. Rather, they address fairness indirectly by highlighting unfair uses of AI systems. Indeed, xAI methods that address bias in this manner also address fairness. However, though the notion of individual fairness is not addressed, there are other forms of fairness that can be addressed by xAI methods.

One such method is the presentation of **ustun_actionable_2019**’s actionable recourse. Suppose some AI system $S$ makes yields some outcome $y$ for some person $P$ represented by feature set $X$. Suppose further that $y$ is an unfavourable outcome for $P$, and $y'$ would have instead been a favourable outcome. Actionable recourse information, in this case, is information that would allow $P$ to change their representation to feature set $X'$ such that $M(X') = y'$. In other words, recourse informs individuals what they would have to change in order to be received more favourably by an AI system. As **ustun_actionable_2019** argue, the mere presentation of this information makes a system more fair, as the decisions made by this system are accompanied by a means of achieving a more favourable decision, so individuals are always given the opportunity to improve their outcomes.

However, while this creates a form of fairness where individuals' outcomes are less fixed, it does not ensure individual fairness. This is by design – actionable recourse does not handle the differential treatment of similar individuals [**ustun_actionable_2019**]. Indeed, one might contend, as **Fleisher_2021** does, that individual fairness is sometimes undesirable, especially when it conflicts with other notions of fairness.

### 2.2.4 Transparency

#### The Problem of Transparency

The problem of transparency is primarily motivated by the desire of stakeholders in an AI system to understand what a given system does and why. For example, in the above COMPAS example, a defendant may desire to understand why their bail was set where it was. In this case, we might consider it insufficient to simply say that COMPAS was used, as we might desire to understand why COMPAS set the recidivism prediction as it is. For most complicated AI systems, this desire cannot be met outright.

An ancillary concern here is trust. While the COMPAS example is not one in which the defendant need trust the system, there are many potential implementations of AI systems where the system's ability to work is contingent on the user trusting the system. For example, consider a system that analyses patients diagnosed with depression and suggests antidepressants as in **jacobs_how_2021**. This work is traditionally done by clinicians, but with the assistance of a machine, the work could potentially be made both easier and more accurate. However, these benefits are contingent on the clinicians' appropriate trust in the system assisting them. That is, they should neither blindly trust nor blindly distrust the machine. Rather, they should trust the machine appropriately, relying upon the machine when it is reliable and ignoring it when it is not. In order to achieve this, transparency into the AI system's reasoning process is desirable.

There are, however, times where total transparency is undesirable. Most obviously, scoring algorithms that are completely transparent to their users can be gamed. However, there are other reasons for algorithms to be less-than-transparent to their users. For example, if a system is optimised around a given condition, explaining this reasoning to a decision-maker may make them more likely to override given decisions, which will only lead to a less optimal system.

### 2.2.5 Addressing Transparency

Unlike bias and fairness, transparency can be directly addressed through explainability methods. Indeed, any explanation of a system or that system's outputs will increase that system's transparency. As such, all explanation will have some impact on transparency. However, this does not imply that explanation constitutes a complete solution to the problem of transparency.

Consider local explanations such as **lundberg_unified_2017**'s SHAP and **ribeiro_anchors_2018**'s Scoped Anchors. Both explanations yield insights into a particular prediction made by a model. However, in their most straightforward application, neither provides transparency into the model in general – rather, they provide transparency into the model in this particular use case. Scoped Anchors, in particular, is designed to give transparency into a single use case – by bounding the given feature with an Anchor, we can replace the model locally with a decision rule [**ribeiro_anchors_2018**]. SHAP, on the other hand, can be extended to provide transparency into a model in general by taking the average feature importances across the distribution of the training set, yielding feature importances for the model overall [**lundberg_unified_2017**].

It is also worth noting that not all explanations yield equivalent forms of transparency. For example, while a question like "why does the model yield a particular prediction in a particular case" is well-answered with Scoped Anchors, it is not well-answered by SHAP [**lundberg__unified__2017**, **ribeiro__anchors__2018**]. Evaluations such as those by **binns__human__2022** and **rader__explanations__2018** indicate that different forms of explanation answer vastly different questions. Thus, the particular problem of transparency in question will have a strong bearing on what explanation method should be used to solve it.

### 2.2.6 Diversity [WIP]

[To-do]...

## 2.3 Prior Applications of IAIDSTs to Diverse, Equitable, and Inclusive Talent Identification [WIP]

We examine other approaches to designing decision support tools for talent identification professionals. We examine computational approaches to fairness here, including the 'similar treatment' principle. We pay special attention to measurements of diversity including the straightforward majority/minority group method and the Entrofy algorithm.

## 2.4 Methodologies [WIP]

### 2.4.1 Research Through Human-Centered Design

This section enumerates research-through-design methods and explores relevant method papers. It focuses on the design methodologies employed in building the SHAP algorithm and the SPF algorithm.

### 2.4.2 Quantitative Methods and Statistical Analysis

This section covers both collecting information from users (i.e., online surveys) and robust statistical analyses applied both to those surveys and to the work in generative AI detection.

### 2.4.3 Qualitative Methods and Thematic Analysis

This section focuses on running interviews and group thinkaloud sessions. It also covers thematic analysis following Braun and Clarke.

# 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 [**binns_human_2022**]. When AI systems make decisions impacting the user, user insight into the decision-making process allows for recourse [**ustun_actionable_2019**]. 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 [**binns_human_2022**, **ustun_actionable_2019**, **wachter_counterfactual_2017**].

The growing field of Explainable Artificial Intelligence (xAI) aims to develop methods for explaining algorithms. While some models are intrinsically interpretable [**rudin_interpretable_2021**], 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 [**molnar_interpretable_2019**]. 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 [**lundberg_unified_2017**, **ribeiro_why_2016**, **ribeiro_anchors_2018**]. 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 [**wang_transparency_2022**].

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 **??**, 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. However, though trust is much-discussed, trustworthiness is often overlooked. Here, we examine literature pertaining to the relationship between trust and trustworthiness, especially in the context of explainable AI. We revisit the framework for trust posited in Chapter* **??***, and discuss the limitations this places on trustworthy explainability systems.*

**jacovi_formalizing_2021** 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_formalizing_2021** 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_how_2021** similarly isolate three elements: "trust is linked to a situation of vulnerability and positive expectations, and is an attitude"; **lee_trust_2004** 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 [**hardin_trust_2002**]. In the machine context, **jacovi_formalizing_2021** 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

*It is important that users be led to place appropriate, calibrated trust in the AI systems they rely on. Exploring the literature, we find that, in many cases, post-hoc explanations of AI outputs increase user trust in the AI system..*

### 3.3.1 The Dangers of Transparency

It is often assumed in the xAI literature that transparency of AI systems is always desirable [**molnar_interpretable_2019**, **miller_explanation_2017-1**]. However, **wang_transparency_2022** 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 [**wang_transparency_2022**].

**ustun__actionable__2019** 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 [**ustun__actionable__2019**].

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 [**ford__play__2020**, **jacobs__how__2021**, **wang__transparency__2022**].

**ford__play__2020** 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 miss-classifications as more correct [**ford__play__2020**].

**jacobs__how__2021** 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_explanation_2017-1** prefer the usage of feature-based explanations, so we find this particularly alarming [**jacobs_how_2021**].

**mccradden_when_2021** responds to **jacobs_how_2021** questioning the role of accuracy in explanations of IAIDSTs. They note the focus of **jacobs_how_2021** 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 [**mccradden_when_2021**].

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_trust_nodate** 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 [**mohseni_trust_nodate**].

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

*Explainable AI (xAI) methods are often motivated by the need to increase user trust in AI systems. However, as we have explored above, increased trust may not always be desirable. While unwarranted distrust might lead people to neglect AI systems when they are correct, unwarranted trust can also be a risk. Explanations which encourage unwarranted 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 SHAP, Scoped Anchors, and a 'Confidence' 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 human-in-the-loop tasks.*

### 3.4.1 The Studies

We design two studies to answer the question: "Can explanations of AI outputs create unwarranted user trust in those outputs?". In particular, we restrict our research question to tasks familiar to lay people with a well-defined but difficult-to-ascertain ground truth. Furthermore, to help make comparisons to related works, we select tasks common in xAI research. Our two tasks are: *estimating a hypothetical person's salary* based on census information of that individual, and *predicting whether someone will be severely delinquent in making a credit payment.* We use two datasets: the Adult dataset collected in the 1994 US Census and curated by **kohavi_scaling_1996** and the Give Me Some Credit dataset curated by **GiveMeSomeCredit**. In both tasks, the participant aims to accurately estimate the dependent variable with the help of the AI system and one of several possible explanations of the AI system's estimate (which is possibly just the confidence rating of the model). More detail on the selection of our tasks can be found in Appendix **??**.

We have preregistered some of the analyses in both of our studies in the OSF registries [**natarajan_binns_2022**]. In particular, we preregistered the analyses in Subsections **??** and **??** for both tasks. We also preregistered the analyses in Subsections **??** and **??**. Finally, we preregistered the analysis in Subsection **??** for the Credit Prediction task only after observing interesting results in the Salary Estimation task.

### 3.4.2 The Model

In both cases, we construct a model using random forests and 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 three explanatory conditions: SHAP, Anchors, and Confidence.

More information on the models themselves and the explanatory conditions is provided in Appendix **??**.

### 3.4.3 Study Design

Both studies rely on the same 3-between-by-2-within design and have the same two factors. The between-subjects 'explanation type' factor determines which explanation a given participant will receive. It is one of 'SHAP', 'Anchor', or 'Confidence'. This between-subjects factor determines which explanation a given participant will receive, and is constant throughout all 6 cases participants are asked to answer. The within-subjects factor is the repeated-measures 'explanation presence' factor. This is either 'before explanation' or 'after explanation'. A flowchart of the study design can be found in Figure **??**.

In both studies, each participant was shown a brief explanation of the task in question and was then asked to complete the 6 cases, with participants given a random mix of correct and incorrect cases. In each case, participants are first

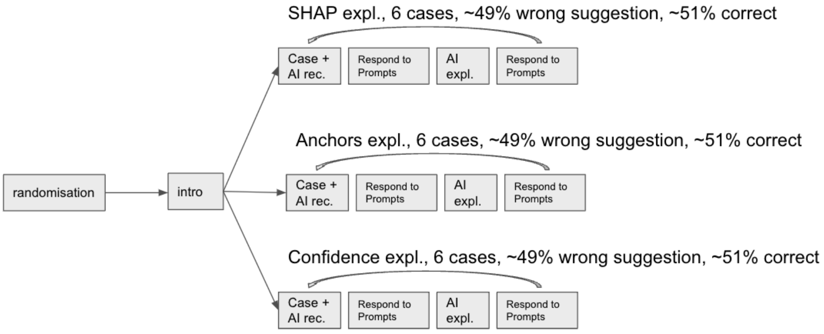


**Figure 3.1:** Study flow

shown a table identifying subject of the case and an AI recommendation of what determination they should make. They are then asked to estimate the dependent variable, and rate both their confidence in the estimate and their trust in the AI recommendation on sliding scales (this is discretised to 20 points).

We code the participant's estimate as a binary *estiamte* variable. The two sliding scale responses are coded as $confidence$ and $trust_{attitude}$ and have values between 1 and 20. As we do this in both the 'before-explanation' and 'after-explanation' conditions, we collect six responses from each participant in each case: $estimate^{before}$, $confidence^{before}$, $trust_{attitude}^{before}$, $estimate^{after}$, $confidence^{before}$, and $trust_{attitude}^{after}$. We additionally have the binary variables *answer* and *recommendation* indicating to true value and AI determination of the dependent variable, respectively.

In addition to these, we define $agreement^{before}$ and $agreement^{after}$ to be whether the user's estimate is in agreement with the machine's recommendation. We define $trust^{b}_{behaviour}efore$ and $trust^{a}_{behaviour}fter$ to be the extent to which the participant's confidence agrees with the machine's recommendation. Finally, in order to reason about the change in a variable due to the explanation, we define '$\Delta$' constructs for all variables with a $before$ and an $after$ case as the after-explanation value minus the before-explanation value.

More detailed definitions of all constructs can be found in Appendix **??**.

### 3.4.4 Demographcs

We had a total of 192 participants complete the Salary Estimation study. These were split pseudorandomly 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 oldest being 18 and the oldest 74. Each applicant completed an introductory page and six cases. The average completion time for these tasks were 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 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 oldest being 20 and the oldest 77. Each applicant completed an introductory page and six cases. The average completion time for these tasks were 7 minutes 53 seconds, the minimum was 2 minutes 17, and the maximum was 30 minutes 13.

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 was marked incomplete and removed from consideration leaving a total of 192 and 197 participants in each study.

### 3.4.5 SHAP and Confidence Increase Unwarranted Trust

We first test for possible increases in unwarranted trust and find that SHAP and Confidence both increase unwarranted trust on our tasks [**natarajan_binns_2022**]. That is, when the AI recommendation is incorrect (I.e., when $answer \neq recommendation$), we find that these conditions increase trust. Furthermore, we find this result more strongly for behavioural trust than attitudinal trust in all but one test. Notably, the Anchors condition does not follow this pattern.

Table **??** shows the results of a one-sided t-test comparing $trust^{after}$ to $trust^{before}$ when $answer \neq recommendation$. A positive $t$ statistic indicates $x^{after} > x^{before}$ (I.e., $\Delta x > 0$), and a negative $t$ statistic indicates $x^{after} < x^{before}$, but, as these are one-sided tests, $p$-values will only be meaningful when $t > 0$. We show both types of trust in all three explanatory conditions on both tasks.

**Table 3.1:** One-Sided T-Tests Comparing Trust Before- and After-Explanation

| Task | Condition | Variable | t Statistic | p Value |
|---|---|---|---|---|
| Salary Estimation | Anchors | $trust_{behaviour}$ | 0.509 | 0.306 |
| | | $trust_{attitude}$ | 0.165 | 0.434 |
| | SHAP | $trust_{behaviour}$ | **3.811** | **< 0.001** |
| | | $trust_{attitude}$ | −0.886 | 0.812 |
| | Confidence | $trust_{behaviour}$ | **2.196** | **0.015** |
| | | $trust_{attitude}$ | 0.945 | 0.173 |
| Credit Prediction | Anchors | $trust_{behaviour}$ | 1.396 | 0.082 |
| | | $trust_{attitude}$ | −2.364 | 0.990 |
| | SHAP | $trust_{behaviour}$ | 1.516 | 0.066 |
| | | $trust_{attitude}$ | **2.475** | **0.007** |
| | Confidence | $trust_{behaviour}$ | **1.835** | **0.034** |
| | | $trust_{attitude}$ | 0.940 | 0.174 |

Note that this is only testing for an unwarranted increase in trust, and says nothing about warranted increases or unwarranted decreases. We instrument this test as such because prior work indicates that modern explanation methods are not well-calibrated, leading us to expect such unwarranted increases [**miller_explainable_2023**].

### 3.4.6 Different Explanation Styles Have Different Effects on Unwarranted Trust

We next seek to determine whether the change in trust varies between our conditions [**natarajan_binns_2022**]. We find that such changes in trust do in fact vary between conditions. Further analysis of this variance is found in Subsections **??** and **??**.

Table **??** contains ANOVA tests examining whether different explanation styles yielded different $\Delta trust$ values when $answer \neq recommendation$. An $F > 1$

indicates that different explanation styles (I.e., SHAP, Anchors, and Confidence) have a different impact on the given trust variable, and $p < 0.05$ indicates that this difference is statistically significant. However, ANOVA analyses do not indicate *which* styles differ.

**Table 3.2:** ANOVAs Comparing Trust Across Explanation Styles

| Task | Variable | F Statistic | p Value |
|---|---|---|---|
| Salary Estimation | $\Delta trust_{behaviour}$ | **3.671** | **0.026** |
| | $\Delta trust_{attitude}$ | 0.925 | 0.397 |
| Credit Prediction | $\Delta trust_{behaviour}$ | 0.066 | 0.936 |
| | $\Delta trust_{attitude}$ | **6.213** | **0.002** |

In the Salary Estimation task, we find no significant results for our ANOVA $trust_{attitude}$, but do find significant results for $trust_{behaviour}$. However, in the Credit Prediction task, we find significant results for our ANOVA $trust_{attitude}$, but none for $trust_{behaviour}$. That is, in the Salary Estiamtion task, different explanations styles have a different impact on behavioural trust, and in the Credit Prediction task, different explanations styles have a different impact on attitudinal trust. We examine these two findings in Subsections **??** and **??**, respectively.

### SHAP Increases Behavioural Trust More than Anchors in the Salary Estimation Task

We would expect that, because we found that $\Delta trust_{behaviour} > 0$ in both the SHAP and Confidence cases of the Salary Estimation Task (and in the Anchors case we did not), the means of the former two should be significantly greater than the latter. However, while we find that SHAP increases behavioural trust more than Anchors, we find no significant results relating to the Confidence condition.

Following our preregistered protocol for significant ANOVA results, we turn to Tukey's HSD as a post-hoc test [**natarajan_binns_2022**]. Table **??** shows the results of this test. Note that this is again restricted to when $answer \neq recommendation$.

**Table 3.3:** Tukey's HSD Test Comparing Change in Behavioural Trust Across Explanations in Salary Estimation

| Condition A | Condition B | Variable | Test Statistic | p Value |
|---|---|---|---|---|
| SHAP | Anchors | $\Delta trust_{behaviour}$ | **2.310** | **0.022** |
| Confidence | Anchors | $\Delta trust_{behaviour}$ | 0.855 | 0.599 |
| SHAP | Confidence | $\Delta trust_{behaviour}$ | 1.455 | 0.198 |

Table **??** demonstrates a significant difference in the mean of $\Delta trust_{behaviour}$ only between the SHAP and Anchors conditions when $answer \neq recommendation$. This indicates that, beyond increasing behavioural trust in incorrect AI outputs, SHAP increases behavioural trust in incorrect AI outputs *more* than Anchors.

### SHAP and Confidence Increase Unwarranted Attitudinal Trust More than Anchors in the Credit Prediction Task

We found a large negative $t$ value for $\Delta trust_{attitude}$ in the Anchors case in the Credit Prediction portion of Table **??** – an effect that is not significant due to the one-sidedness of our tests. However, as we found only positive $t$ values for $\Delta trust_{attitude}$ in the SHAP and Confidence cases, we expect that Anchors has a negative effect on $\Delta trust_{attitude}$ relative to SHAP and Confidence. Table **??** confirms this expectation: SHAP and Confidence both have a more positive effect on attitudinal trust than Anchors in the Credit Prediction task.

Table **??** again follows our preregistered protocol for significant ANOVA results with a Tukey's Honestly Significant Difference (HSD) test [**natarajan_binns_2022**]. This is again restricted to when $answer \neq recommendation$.

**Table 3.4:** Tukey's HSD Test Comparing Change in Attitudinal Trust Across Explanations in Credit Prediction

| Condition A | Condition B | Variable | Test Statistic | p Value |
|---|---|---|---|---|
| SHAP | Anchors | $\Delta trust_{attitude}$ | **1.213** | $<$ **0.001** |
| Confidence | Anchors | $\Delta trust_{attitude}$ | **1.030** | $<$ **0.001** |
| SHAP | Confidence | $\Delta trust_{attitude}$ | 0.183 | 0.708 |

Table **??** shows a significant difference in the mean of $\Delta trust_{behaviour}$ between the Anchors condition and both other conditions, but no significant difference between SHAP and Confidence. This indicates that SHAP and Confidence increase unwarranted attitudinal trust relative to Anchors (although Anchors appears to actually *reduce* unwarranted attitudinal trust).

Note that this does not prove that Anchors reduces attitudinal trust relative to no explanation. For this analysis, we will need another t-test. As we did not preregister this test, analysis of this phenomenon is included in exploratory analysis in Subsection **??** [**natarajan_binns_2022**].

### 3.4.7 Anchors Decrease Attitudinal Trust in a Machine

We noted already that SHAP and Confidence appear to increase trust in cases where the machine is incorrect. However, we noticed no such result for Anchors. Instead, we found that Anchors appeared to decrease end-user trust in incorrect machine outputs (I.e., when $answer \neq recommendation$). The two-sided t-test in table **??** confirms this.[1]

**Table 3.5:** Two-Sided T-Tests Comparing Trust Before- and After-Explanation

| Task | Condition | Variable | t Statistic | p Value |
|---|---|---|---|---|
| Salary Estimation | Anchors | $trust_{behaviour}$ | 0.509 | 0.611 |
| | | $trust_{attitude}$ | 0.165 | 0.869 |
| Credit Prediction | Anchors | $trust_{behaviour}$ | 1.396 | 0.164 |
| | | $trust_{attitude}$ | $\mathbf{-2.364}$ | **0.019** |

Table **??** shows that, on the two-sided t-test, we *do* find that the provision of Anchors explanations decreases participant attitudinal trust in incorrect machine recommendations, at least in the Credit Prediction task. However, we do not see a similar effect on behavioural trust, and this effect is limited to only one task.

[1] This analysis was not preregistered.

### 3.4.8 Behavioural and Attitudinal Trust are Highly Correlated

Given that many patterns observed for $trust_{behaviour}$ do not hold for $trust_{attitude}$, we might expect these variables to correlate only minimally. However, while they are mathematically distinct constructs, they are both intended to measure the same underlying phenomenon. Table **??** demonstrates a high correlation between these two measurements of trust across all cases.[2][3]

Table **??** shows a Pearson's correlation analysis across all explanatory conditions in both the before- and after- cases. We also perform this analysis on $\Delta trust_{attitude}$ and $\Delta trust_{behaviour}$.

**Table 3.6:** Pearson's Correlation Between Attitudinal and Behavioural Trust

| Task | Variable A | Variable B | Rho | p Value |
|---|---|---|---|---|
| Salary Estimation | $trust_{attitude}$ | $trust_{behaviour}$ | **0.630** | $< \mathbf{0.001}$ |
| | $\Delta trust_{attitude}$ | $\Delta trust_{behaviour}$ | **0.265** | $< \mathbf{0.001}$ |
| Credit Prediction | $trust_{attitude}$ | $trust_{behaviour}$ | **0.612** | $< \mathbf{0.001}$ |
| | $\Delta trust_{attitude}$ | $\Delta trust_{behaviour}$ | **0.179** | $< \mathbf{0.001}$ |

Table **??** indicates that both trust variables are highly correlated across both of our tasks. Furthermore, though the correlation between the $\Delta trust$ is more modest, it is still statistically significant.

### 3.4.9 Anchors and SHAP Increase Participant Confidence in their Prediction

Noting that behavioural trust is constructed from $confidence$ values, we ask: "does providing an Anchors explanations increase participant confidence in their own decisions when the AI recommendation is incorrect?" Table **??** supplies evidence

[2] Though previous analyses considered only incorrect recommendations, this analysis relates *all* trust, not just unwarranted trust, so we consider all cases, regardless of whether or not $answer == recommendation$.

[3] This analysis was only partially preregistered; we did not register this analysis in the Salary Estimation task, but we did in the Credit Prediction task [**natarajan_binns_2022**].

indicating that Anchors (and SHAP) explanations increase participant confidence in their own estimations when $answer \neq recommendation$. This result agrees with **wan_explainabilitys_2022** and **bansal_does_2021**, suggesting these explanation types yield a blanket increase in participant self-confidence.[4]

Table **??** contains a t-test comparing $confidence$ in all three explanatory conditions. [5]

**Table 3.7:** One-Sided T-Tests Comparing $confidence$ Before- and After-Explanation

| Task | Condition | Variable | t Statistic | p Value |
|---|---|---|---|---|
| Salary Estimation | Anchors | $confidence$ | **2.171** | **0.016** |
| | SHAP | $confidence$ | **1.694** | **0.046** |
| | Confidence | $confidence$ | 1.047 | 0.296 |
| Credit Prediction | Anchors | $confidence$ | **1.742** | **0.042** |
| | SHAP | $confidence$ | **3.473** | $< \mathbf{0.001}$ |
| | Confidence | $confidence$ | 0.752 | 0.226 |

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

### 3.4.10 Explanations Impact Trust Differently When the Machine is Correct

Note that properly calibrated trust would involve both distrusting the machine when it is wrong and trusting it when it is right. To assess the latter, we give a brief evaluation of what happens in the cases where the AI is correct, I.e. where $answer == recommendation$. Table **??** contains the results of these analyses.[6]

[4] This analysis was not preregistered.

[5] Note for clarity that $confidence$ is the variable indicating participant confidence in their own decisions, and 'Confidence' is the condition in which the machine's explanation consists of its own confidence in its suggestion.

[6] These analyses were not preregistered.

**Table 3.8:** Two-Sided T-Tests Comparing Trust Before- and After-Explanation when *answer == recommendation*

| Task | Condition | Variable | F Statistic | p Value |
|---|---|---|---|---|
| Salary Estimation | Anchors | $trust_{behaviour}$ | 0.502 | 0.616 |
| | | $trust_{attitude}$ | **−2.337** | **0.020** |
| | SHAP | $trust_{behaviour}$ | 0.295 | 0.768 |
| | | $trust_{attitude}$ | −1.385 | 0.168 |
| | Confidence | $trust_{behaviour}$ | **2.410** | **0.017** |
| | | $trust_{attitude}$ | **3.254** | **0.001** |
| Credit Prediction | Anchors | $trust_{behaviour}$ | **3.013** | **0.003** |
| | | $trust_{attitude}$ | **−2.487** | **0.014** |
| | SHAP | $trust_{behaviour}$ | 0.207 | 0.836 |
| | | $trust_{attitude}$ | **3.538** | **0.001** |
| | Confidence | $trust_{behaviour}$ | **2.863** | **0.005** |
| | | $trust_{attitude}$ | **2.461** | **0.015** |

### Confidence Explanations Increase Warranted Trust

Positive $\Delta trust$ values in all cases in table **??** demonstrate that Confidence explanations increased both measured types of user trust in the AI output when the model is correct across both tasks.

### Anchors Explanations Decrease Warranted Attitudinal Trust

Table **??** also demonstrates that providing Anchors explanations yields a *decrease* in $trust_{attitude}$. This, alongside the finding from Subsection **??**, indicates that Anchors explanations have a negative impact on $trust_{attitude}$, regardless of model correctness.

### SHAP Explanations Increase Warranted Attitudinal Trust in the Credit Prediction Task

SHAP explanations increase Warranted attitudinal trust in the Credit Prediction Task as shown in table **??**. However, the effect is not mirrored in tests of $trust_{behaviour}$ or on the Salary Estimation Task.

### 3.4.11 Limitations

#### Generalisation and External Validity

The external validity of our results may be challenged due to our use of artificial tasks and benchmark datasets. Indeed, field studies with decision-makers in real deployments would be needed to yield results that apply uncontroversially in a given domain. Along a similar vein, one might contend that the use of only two tasks is insufficient to generalise our results to other domains. One might also challenge external validity of our results due to our limited selection of only three explanatory conditions (only two of which are commonly used xAI methods).

We choose two similar datasets in a narrow domain (human-in-the-loop binary classification tasks with definite but non-obvious ground truth) as we do not wish to confuse our primary research questions with questions surrounding generalisation. Similarly, we contend research on further tasks is extraneous to the primary questions. More information on our specific selection of cases can be found in Appendix **??**.

We choose SHAP and Anchors as popular candidate explanations from different styles, as we wish to demonstrate conditions under which unwarranted trust may and may not arise. We choose Confidence as a third condition as this condition acts as a sort of baseline. More information on our choice of explanation algorithms can be found in Appendix **??**.

That said, insofar as performance on benchmark datasets can be expected to generalise, we believe our findings will extend to predictive tasks whose answers are difficult enough to both human and AI systems and where performance is distributed differently between them. We do not expect our findings to generalise to tasks where the answers are either immediately evident to the user (such as predicting whether a given image contains a cat) or so difficult as to be impossible to the naked eye (such as gene function prediction). And while we do expect our findings to generalise beyond tabulated binary classification tasks, said generalisation does not impact the primary significance of these findings: xAI developers, evaluators, and implementers alike should concern themselves with the possibility of unwarranted trust.

#### Effect on Overall Task Performance

We have focused exclusively on the phenomenon of misplaced trust when the AI output is incorrect. However, while problematic, this has to be weighed against the effect an xAI method has on overall task performance. It might be remarked that, even if SHAP explanations increased misplaced trust, if they increased performance on the task overall, that would still be desirable. Indeed, relative to no explanation at all, SHAP may be beneficial simply because of its effect of increasing user trust overall. (The effect of an overall increase in trust in AI outputs will depend on the relative accuracies of the AI and the human.)

However, we do not contend that the unwarranted trust issue renders SHAP explanations useless. We instead contend that they may lead to dangerous abuse. When implementing xAI algorithms in human-in-the-loop tasks, implementers should consider the possible harms of this potential for abuse, especially when these tasks have definite but unobvious ground truth. Furthermore, those developing xAI applications for use in these contexts should strive to develop explanation methods that appropriately calibrate trust in addition to increasing overall task performance.

### 3.4.12 Discussion

In this chapter, we investigate the extent to which three different AI explanation methods can create unwarranted trust in their underlying AI systems on a specific set of tasks. Specifically, we look at whether any of our Anchors, SHAP, or Confidence explanatory conditions increase trust in a model when that model is wrong. We restrict our analysis specifically to the subset of tasks human-in-the-loop where the ground truth of the prediction is neither subjective nor immediately evident to the human. We perform analyses on two tabular binary prediction tasks that satisfy this condition: Salary Estimation and Credit Prediction.

We conclude here that SHAP and Confidence explanations are liable to induce unwarranted trust under these circumstances, but we find no evidence of the same for Anchors. This demonstrates that, in these cases, neither SHAP nor Confidence serve to correctly calibrate trust in AI outputs. Rather, they blindly increase trust in these outputs, encouraging users to incorrectly agree with the AI explained. Furthermore, we find that, while Anchors may not induce unwarranted trust, it instead fosters blanket distrust for the AI and increased confidence in user decision-making. As such, the effect of Anchors explanations on trust calibration may not be positive overall; it is dependent on the distribution of correct vs. incorrect disagreements between user and AI. (Indeed, for certain distributions of correct vs. incorrect disagreements, all three explanation styles may have a negative overall effect on trust calibration.) These findings suggest that, at least under the conditions outlined, xAI may be at best useless and at worst actively harmful in its effects on human trust calibration.

Though trust is often a component of both design and evaluation, *trustworthiness* is often overlooked [**jacovi_formalizing_2021**, **lundberg_unified_2017**, **ribeiro_why_2016**, **jacobs_how_2021**]. It is often assumed that these explanations draw directly from model outputs, and are therefore akin to oracles that grant a general understanding of the system. In **natarajan_trust_2023**, we argue for a paradigm shift against this use of explainable AI systems, and instead argue that systems should be developed, evaluated, and used for specific purposes. We agree with this sentiment. In fact, we contend it is not itself problematic that user trust in AI systems might increase as a result of xAI. But it is problematic that user trust in wrong decisions made by AI systems may increase in response to the use of a maximally trust-inducing explanation method. Particularly in human-in-the-loop tasks with definite but unclear ground truth, trust should not be considered the *raison d'etre* of xAI. More heed should be instead paid to *calibrating* trust in AI systems and to ensuring that explanations are sufficiently well-calibrated for the use cases they are considered for.

## 3.5 Conclusion

# 4

# Using Post-Hoc Explainable AI to Identify Talent: Explaining an Applicant Scoring Algorithm with Shapley Values

## Contents

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.

## 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

(To-do)

# 5
# A Human-Centric Approach to Identifying Talent With Intrinsically Interpretable Models [WIP]

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

The SPF (to be elaborated on below) 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

# A Possibility Frontier Approach to Diverse Talent Selection

## Contents

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.

## 6.1 The Diverse Talent Selection Problem

(To-do)

## 6.2 The Mathematics of the Selection Possibility Frontier

(To-do)

## 6.3 A Case Study

(To-do)

# 7

# Student Generative AI Usage in Application Essays [WIP]

Contents

## 7.1 Introduction

To-do

## 7.2 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)

# 8 Discussion

## Contents

## 8.1 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.

## 8.2 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.

## 8.3 Conclusion

To-do

# Appendices

# A
# Definitions and Terminology

**Contents**

## A.1 To-do

To-do

# B

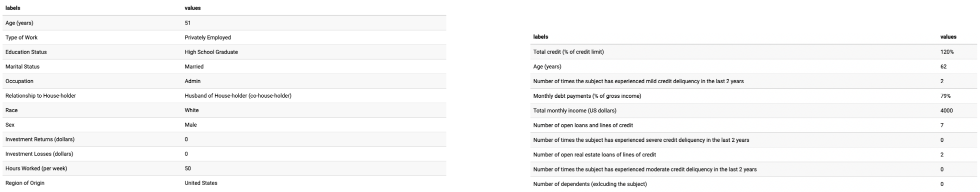
# Reference Materials for Misleading Explanations of AI Outputs in Talent Identification

## Contents

## B.1 Tasks

Rather than an evaluation of a machine (where a user might rate a model as being correct because it is consistent), we restrict our analysis to human-in-the-loop tasks (where, if a user disagrees with a machine, they might override it) to better simulate and assess the sense in which a human-in-the-loop is *vulnerable* to being mislead by an explanation.

Furthermore, we recognize that in many cases, a user's confidence in their own judgement would crowd out any possible effect of the AI system (and any associated explanation), or the ambiguity about whether any ground truth exists at all may lead a user to hold steadfast in their opinions. Thus, we avoid domains like “is this an image of a cat?” where it is obvious to the human whether the image is one of a cat. Similarly, we avoid tasks without a clear ground truth such as “is this language offensive?” or “is this a good work of art?”. (Philosophers continue to argue whether or not these questions have objective answers. Kant, for example, contended that, while they do not admit objective answers, subjects are nonetheless justified in making universal claims [**zangwill__aesthetic__2023**].)



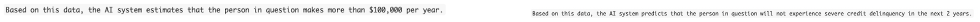
| labels | values |
|---|---|
| Age (years) | 51 |
| Type of Work | Privately Employed |
| Education Status | High School Graduate |
| Marital Status | Married |
| Occupation | Admin |
| Relationship to House-holder | Husband of House-holder (co-house-holder) |
| Race | White |
| Sex | Male |
| Investment Returns (dollars) | 0 |
| Investment Losses (dollars) | 0 |
| Hours Worked (per week) | 50 |
| Region of Origin | United States |

**(a)**

| labels | values |
|---|---|
| Total credit (% of credit limit) | 120% |
| Age (years) | 62 |
| Number of times the subject has experienced mild credit deliquency in the last 2 years | 2 |
| Monthly debt payments (% of gross income) | 79% |
| Total monthly income (US dollars) | 4000 |
| Number of open loans and lines of credit | 7 |
| Number of times the subject has experienced severe credit deliquency in the last 2 years | 0 |
| Number of open real estate loans of lines of credit | 2 |
| Number of times the subject has experienced moderate credit deliquency in the last 2 years | 0 |
| Number of dependents (exlcuding the subject) | 0 |

**(b)**

**Figure B.1:** Features from a Sample Case in (a) the Salary Estimation and (b) the Credit Prediction Tasks



Based on this data, the AI system estimates that the person in question makes more than $100,000 per year.

**(a)**

Based on this data, the AI system predicts that the person in question will not experience severe credit delinquency in the next 2 years.

**(b)**

**Figure B.2:** Predictions from a Sample Case in (a) the Salary Estimation and (b) the Credit Prediction Tasks

We choose two standard benchmark datasets: Adult and Give Me Some Credit [**kohavi__scaling__1996**, **GiveMeSomeCredit**]. We choose these because such datasets are typically the grounds on which xAI methods are developed and critiqued. The Adult dataset specifically has been used as a benchmark in multiple xAI studies, including **weerts__case-based__2019**, **weerts__human-grounded__2019**, and **ribeiro__anchors__2018**. Various credit datasets, including Give Me Some Credit, are common as benchmarks in xAI studies like **bansal__does__2021**, **ustun__learning__nodate**, and **krishna__disagreement__2022**.

The dependent variable (to be estimated by the participant) of the Adult dataset is “Does this person make more than $50,000?”. However, we adjusted the amount due to inflation (this amount in 1994 is roughly $100,000 today). Thus, we ask

“Does this person make more than $100,000?”. The dependent variable of the Give Me Some Credit dataset is “Will this person be at least 90 days delinquent on a credit payment in the next two years”, which we simplify to “Will this person experience severe credit delinquency in the next two years” (we also include a definition of “severe credit delinquency”).

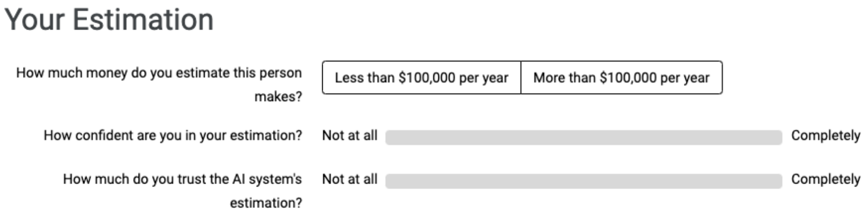
The participant has access to a table containing all data the model does in both cases; samples of these tables are shown in Figure **??**. Figure **??** displays the initial estimate of the AI system for the very same samples, as they are shown to participants.

## B.2 Questions and Constructs

We ask three questions both before and after explanation. In the Salary Estimation task, we ask “How much money do you estimate this person makes?”, to which the possible responses are “Less than $100,000 per year” and “More than $100,000 per year”. In the Credit Prediction task, we ask “Do you predict that this person will experience severe credit delinquency?”, where the answers are “Will not experience severe delinquency” and “Will experience severe delinquency”. In either case, this yields a binary *estimate* variable.

Then, we ask “How confident are you in your estimation?”, which has a 20-point sliding scale of possible responses. This yields the $confidence$ (confidence) variable with a value between 1 and 20.

Finally, we ask “How much do you trust the AI system’s estimation?”, which has a 20-point sliding scale of possible responses. This yields the attitudinal trust variable, $trust_{attitude}$, also with a value between 1 and 20.



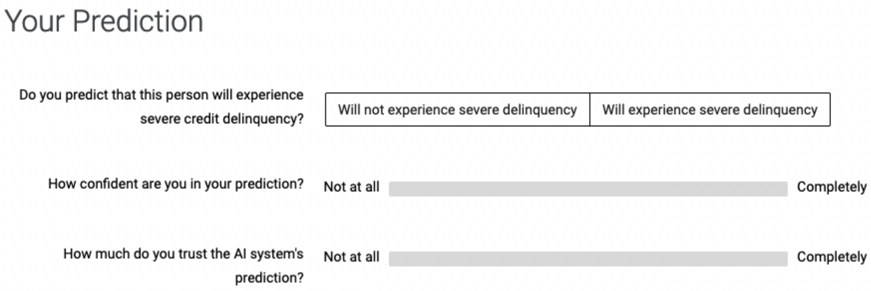
Your Estimation

How much money do you estimate this person makes? Less than $100,000 per year | More than $100,000 per year

How confident are you in your estimation? Not at all Completely

How much do you trust the AI system's estimation? Not at all Completely

**Figure B.3:** Three Questions Asked Twice per Case (Salary Estimation)



Your Prediction

Do you predict that this person will experience severe credit delinquency?

How confident are you in your prediction?

How much do you trust the AI system's prediction?

Will not experience severe delinquency
Will experience severe delinquency
Not at all
Completely
Not at all
Completely

**Figure B.4:** Three Questions Asked Twice per Case (Credit Prediction)

These three questions are asked twice; once before the participant sees one of the three experimental conditions (SHAP, Anchors, and Confidence), and once again after. We index the 'before' case as $before$ and the 'after' case as $after$.

Thus, we collect six variables from participants per case: $estimate^{before}$, $confidence^{before}$, $trust^{before}_{attitude}$, $estimate^{after}$, $confidence^{after}$, $trust^{after}_{attitude}$. The presentation of the three questions is shown in Figures **??** and **??**.

We also collect two additional variables per case: $answer$ and $recommendation$. In each case, $answer$ is the correct output for the case (as in our test output data), and $recommendation$ is the machine's recommendation of what the user should estimate for that case.

We can now define some constructed variables for use in our analysis. First, we define a $agreement$ to be whether the user's estimate is in agreement with the machine's recommendation in Equation **??**.

$$agreement^x := estimate^x == recommendation \tag{B.1}$$

We now define a variable $trust_{behaviour}$ in Equation **??**, which is a combination of the $estimate$, $confidence$, and $agreement$ variables to yield a 40-point scale of behavioural trust. Here, $-19$ is absolute confidence that the system is wrong and 20 is absolute confidence that the system is right. Let:

$$trust^x_{behaviour} := \begin{cases} confidence^x & agreement^x \\ 1 - confidence^x & otherwise \end{cases} \tag{B.2}$$

In order to reason about the change in a variable due to the explanation, we define '$\Delta$' constructs as the difference between after- and before-explanation for each variable, as seen in Equation **??**.

$$\Delta x := x^{after} - x^{before} \tag{B.3}$$

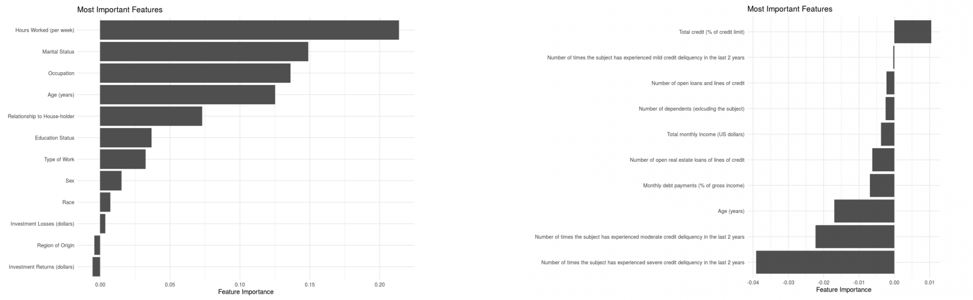
Informally, $\Delta trust_{attitude}$ is the effect of the explanation on the participant's attitudinal trust in the AI determination. Formally, $\Delta trust_{attitude} := trust^{after}_{attitude} - trust^{before}_{attitude}$. E.g., suppose a participant sees a determination they believe to be incorrect in the absence of an explanation, they might express distrust in that system and rate their trust low (say 3); suppose further that, the explanation provided is persuasive, and the participant rates their trust high (say 19) after reading it; the $\Delta trust_{attitude}$ for this case would be $19 - 3 = 16$.

## B.3 Models

In this study, we rely on three different models. We use a random forest classifier as our base model. We use a SHAP explainer to produce one of our explanatory conditions, and a Scoped Anchors explainer to produce another (our final explanatory condition is produced natively by the random forest).

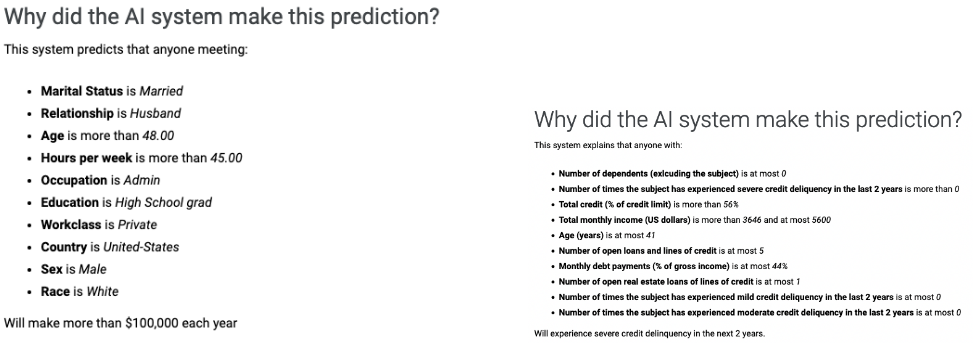
We opt for a random forest model as they are still ubiquitous in the field of AI, and the Adult and Give Me Some Credit datasets lack the kind of feature richness required to noticeably benefit from more sophisticated model architectures [**Grinsztajn_Oyallon_Varoquaux_2022**]. The random forest classifiers classify points in each data set into either of the two outcome conditions. Our random forest classifier achieves 86% test accuracy on the Adult dataset and 93% test accuracy on the Give Me Some Credit dataset.

Our three explanatory conditions are: SHAP, Anchors, and Confidence. SHAP explanations show participants a plot of feature importances, where the presented importances are the Shapley Values calculated for each feature. In the Anchors explanations, participants are shown rules that nearly guarantee that other cases



(a) (b)

**Figure B.5:** Sample SHAP Explanations in (a) the Salary Estimation and (b) the Credit Prediction Tasks



Why did the AI system make this prediction?

This system predicts that anyone meeting:

- **Marital Status** is *Married*
- **Relationship** is *Husband*
- **Age** is more than *48.00*
- **Hours per week** is more than *45.00*
- **Occupation** is *Admin*
- **Education** is *High School grad*
- **Workclass** is *Private*
- **Country** is *United-States*
- **Sex** is *Male*
- **Race** is *White*

Will make more than $100,000 each year

(a)

Why did the AI system make this prediction?

This system explains that anyone with:

- **Number of dependents (exlcuding the subject)** is at most *0*
- **Number of times the subject has experienced severe credit deliquency in the last 2 years** is more than *0*
- **Total credit (% of credit limit)** is more than *56%*
- **Total monthly income (US dollars)** is more than *3646* and at most *5600*
- **Age (years)** is at most *41*
- **Number of open loans and lines of credit** is at most *5*
- **Monthly debt payments (% of gross income)** is at most *44%*
- **Number of open real estate loans of lines of credit** is at most *1*
- **Number of times the subject has experienced mild credit deliquency in the last 2 years** is at most *0*
- **Number of times the subject has experienced moderate credit deliquency in the last 2 years** is at most *0*

Will experience severe credit delinquency in the next 2 years.

(b)

**Figure B.6:** Sample Anchors Explanations in (a) the Salary Estimation and (b) the Credit Prediction Tasks

following these rules will have the same estimate. In the Confidence condition, participants are shown the model's confidence in its estimate (based on the average output of all decision trees in the random forest model). We expect that providing any of these three explanations will generate an increase in trust even when the AI system is incorrect.

We select these three conditions carefully. We wish to test the popular feature importance explanation methods, of which SHapley-based Additive exPlanations, or

The system is 82% sure of its estimate

(a)

The system is 76% sure of its prediction

(b)

**Figure B.7:** Sample Confidence Explanations in (a) the Salary Estimation and (b) the Credit Prediction Tasks

SHAP (the 'SHAP' condition), is a prominent example [**lundberg_unified_2017**]. These explanations weight the importances of different features in a particular prediction. They are are often used to highlight aspects of the model to users and have been criticised in **kumar_problems_2020** for failing to follow norms of explanations drawn from philosophy, psychology, and cognitive science, making them bad candidates for explanation to end users. We also test the Scoped Anchors explanation algorithm (the 'Anchors' condition) [**ribeiro_anchors_2018**]. These explanations justify individual model predictions following human-centred norms by providing rules that bound model behaviour. **bansal_does_2021** and **jacobs_how_2021** both note that this styles of explanation raises user trust in a model. Finally, we also test the provision of the model's own confidence in its estimate (the 'Confidence' condition). While confidence statistics are not generally regarded as an xAI method, they serve a similar function in that they indicate to the user something about the model's operation which allows evaluation of the model on those grounds.

We fine-tuned the visual design of each condition with a pilot study using the Adult dataset. In the pilot, we asked participants to indicate which of a range of visualisations is the most intuitive, and asked various questions to test their comprehension. For the SHAP condition, we found that a bar plot plotting feature importances, a format used for all feature importance methods across the field, appeared to be considered more intuitive [**weerts_human-grounded_2019**]. As both models have relatively few features, we opted to show all features. For the Anchors condition, we found that a verbal explanation of the Anchors, mimicking formats used by popular social media platforms, was considered the most intuitive [**ribeiro_anchors_2018**]. The formats used in the study itself can be seen in Figures **??**, **??**, and **??**.

## B.4 The Study

We collected data in surveys designed and hosted on Formr.[1] The participants were recruited via Prolific Academic's standard sampling method restricted to the United States (to match the origin of both datasets).[2]

Participants completed the study 6 times with 6 cases drawn from a pool of 39, where the AI system is correct on 20 out of the 39. To avoid learned effects related to the accuracy of either the model or the sample pool, participants were not told the machine accuracy or the distribution of the case pool. No participants were permitted to take part in both studies.

[1] www.formr.org

[2] www.prolific.co

# C
# ChatGPT Generation

**Contents**

## C.1 To-do

To-do