



# The perception of fairness of algorithms and proxy information in financial services

---

A report for the Centre for Data Ethics and Innovation  
from the Behavioural Insights Team

October 2019

# Contents

---

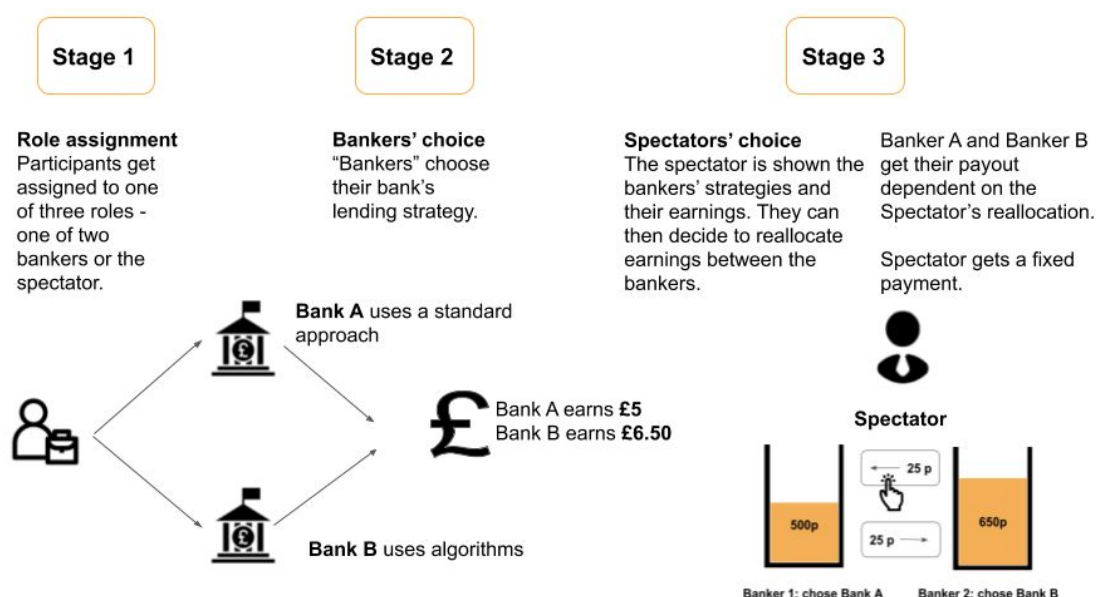
<b>Contents</b>	<b>2</b>
0.1 Executive summary	3
<b>Part 1: Background and trial design</b>	<b>9</b>
1.1 Background	9
1.2 Project aims	10
1.3 Online experiments and Predictiv	10
1.4 Experimental design and procedures	11
1.5 Treatments and randomisation	17
1.6 Participant pool and eligibility	19
1.7 Outcome measures	19
1.8 Analysis strategy	23
1.9 Power calculations	25
1.10 Risks	26
<b>Part 2: Findings and recommendations</b>	<b>27</b>
2.1 Implementation	27
2.2 Results	27
2.3 Conclusion	39
2.4 Recommendations	40
<b>Appendices</b>	<b>42</b>

## 0.1 Executive summary

The Centre for Data Ethics and Innovation (CDEI) was created in late 2018 to connect policymakers, industry, civil society, and the public to develop the right governance regime for data-driven technologies. As part of this mandate, the CDEI is interested in building an evidence base on people's perceptions of fairness of data use in business decision-making, especially in the financial services sector. The Behavioural Insights Team (BIT) and the CDEI collaborated to investigate how people respond to the use of algorithms in a particularly consequential area of everyday life: personal finances. This brings together the CDEI's expertise in algorithmic decision-making and BIT's expertise in human behaviour and robust evaluation. The experiment explores how *fair* people perceive the use of algorithmic decision-making by banks to be when making a decision about loan eligibility. Further it considers whether fairness perceptions vary dependent on the *type* of information the algorithm takes into account.

### Experimental design

The experiment involves an adaptation of an established fairness experimental design, in which two participants act as "bankers", and a third acts as an independent "spectator". We ask a spectator to redistribute earnings between two different bankers. This decision is consequential as it determines actual financial outcomes for the two bankers. We assess how much money the spectator is willing to redistribute in order to punish someone for (perceived) unfair behaviour. This is our key outcome measure capturing perceptions of fairness. We compare how much the spectator reallocates depending on the descriptions of the bankers' strategies. Table 1 describes the four different conditions.



The experiment was conducted online via BIT's [Predictiv platform](#). The bankers were informed that they were senior executives and had to choose between one of two bank strategies: Bank A or Bank B. As part of the experiment, we varied the description of the strategies for Bank B. In the control condition, both options offered a neutrally framed operating model of a bank. It read:

- Bank A uses financial information to determine an individual's application, such as a person's salary, whether someone is employed and if they have debt.
- Bank B uses advanced computing techniques and a broader range of personal information than Bank A to make decisions about loan applications. This allows it to make predictions about an individual's application.

In the three treatment conditions, the description of Bank B varied such that it emphasised the use of different data in their algorithms. It highlighted that Bank B uses information which could act as a proxy for other characteristics, specifically gender,<sup>1</sup> ethnicity or social media usage. The descriptions read as follows:

**Table 1: Descriptions used for Bank B across the four conditions**

Condition	Description
<b>Control (neutral)</b>	Bank B uses advanced computing techniques and a broader range of personal information than Bank A to make decisions about loan applications. This allows it to make predictions about an individual's application.
<b>Gender</b>	[As in control condition, plus]... However, by including information which may link to gender, for example, salary and occupation, Bank B may end up offering different levels of credit to men and women.
<b>Ethnicity</b>	[As in control condition, plus]... However, by including information which may link to ethnicity, for example, salary, postcode and occupation, Bank B may end up offering different levels of credit to people of different ethnicities.
<b>Social media</b>	[As in control condition, plus]...However, by including social media data, Bank B may end up offering different levels of credit to people with a greater social media presence compared to people with a smaller social media presence.

**Across all conditions, choosing Bank A initially gives the banker £5; choosing Bank B initially gives the banker £6.50.** The bankers are asked to make a total of four decisions, one for each of the four algorithm scenarios (which are shown in a random order).

<sup>1</sup> In this research we inadvertently use the term gender rather than sex and will ensure to remedy this in all future research. [The Equality Act](#) makes it illegal to discriminate against someone on the basis of their sex or gender reassignment, but not gender.

**The third participant, the spectator, is informed about each banker's decision, and given the opportunity to reallocate the profits received from one banker to the other.**

The choice of the spectator represents whether they think the allocation is fair and by how much they care. We compare allocation choices across the four different conditions to determine which bank strategy spectators believe is most unfair. For example, it may be that spectators allocate significantly more money away from Bank B, the bank that uses algorithmic decision-making, to Bank A in the ethnicity condition compared to the social media condition.

Finally, in our exploratory research we benchmark reallocation to three different scenarios. This allows us to put our results into perspective by 1) comparing them against a well-known practice that is typically perceived as unethical, and 2) considering how people's existing perceptions of algorithms affect their decision. The scenarios were as follows:

1. Spectators are asked to make a reallocation decision for a scenario where Bank B is registered in the Cayman Islands and does not pay any corporation tax to the UK government. This is a well-known practice that people tend to be averse to.
2. Spectators are asked how accurate they believe Bank A is compared to Bank B in its assessments about whether to give someone a loan or not.
3. Spectators are asked to assume that Bank B is more accurate than Bank A and are asked if they want to change their original reallocation decision (this was hypothetical).

## Findings and implications

### What have we learned from this experiment?

On average:

- People have a negative perception of algorithmic decision-making in loan decisions, considering that 17% of those in the control condition reallocated above the amount needed to equalise returns between the banks.
- People perceive the use of information that could be used as a proxy in loan decisions as unfair.
- Those most at risk of being discriminated against feel most strongly that it is unfair (noting that this is not conclusive).
- People perceive the fairness of this proxy information use roughly a third as much as they do tax avoidance.
- When people think that the algorithm is more accurate, they see it as fairer. This demonstrates that accuracy is of particular value to people.

**On average, 15.5% more people punish a bank financially when informed that it uses information which could act as a proxy for other characteristics (gender, ethnicity or social media usage) in their algorithms compared to a neutral description.** What this means is that a larger proportion of people choose to move more than 75p from Bank A to Bank B in the treatment conditions compared to the control. This is relatively consistent across the treatment conditions, such that 14.3% more do so in the gender condition, 16% in

the ethnicity condition and 16.3% in the social media condition. In the real world, this could result in lower take-up of financial products, consumers switching to other providers or sharing unfavourable reviews about the bank, for example. However, switching rates in general are known to be low,<sup>2</sup> so consumers might not get around to acting even if they are unhappy. In addition, financial providers may be (unintentionally) incentivised not to be transparent about their practices to avoid any risk of losing business.

**Table 2: Breakdown of how people chose to reallocate money across control and treatment**

	Control	Treatments
<b>No action</b>	<b>21.3%</b>	<b>18.1%</b>
<b>Money towards Bank A</b>	<b>50.2%</b>	<b>59.5%</b>
Less than 75p	14.0%	11.9%
Equalising pay (75p)	18.6%	14.7%
More than 75p	17.5%	33.0%
<b>Money towards proxy</b>	<b>28.5%</b>	<b>22.3%</b>

Table 2 describes the different categories of reallocation by condition. **Spectators could reallocate up to £6.50 from Bank B to Bank A, or up to £5 from Bank A to Bank B. In the control condition, people on average reallocate 28p to Bank A.** This may be driven by factors such as inequity aversion and how they want two quite similar banks to be treated similarly or perhaps aversion to algorithmic decision-making. **This increases considerably to 63p in the gender condition, 64p in the ethnicity condition and 73p in the social media condition, when the use of information which could act as a proxy is introduced.** People's views of social media use in algorithmic decision-making is particularly interesting. Although the differences between conditions is not statistically significant, directionally the results suggest that people feel most uncomfortable with the potential use of social media as a decision metric.

We are interested in whether different groups respond differently to the various conditions, such that whether those most likely perceive themselves to be negatively affected by the information proxies respond more strongly. While we did not have sufficient statistical power to robustly infer causal findings, we see the following:

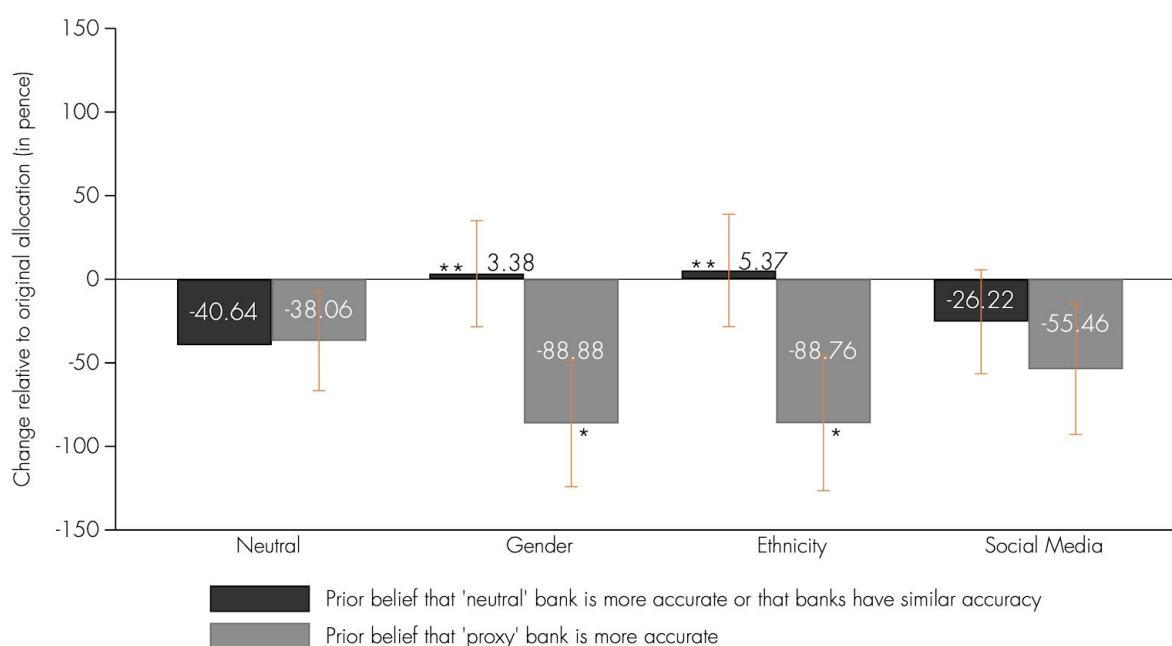
- Directionally, women punish Bank B more strongly than men in the gender condition;
- Overall, individuals who are white seem to punish slightly less compared to ethnic minorities, though the difference is minimal in the ethnicity prompt condition;
- A mixed picture is found in the social media condition such that neither frequency of use or age is related to a stronger response.

<sup>2</sup> See Citizens Advice research on the [Loyalty Penalty](#).

**We see that people punish tax avoidance by reallocating 155p on average to the bank that does not avoid tax payment compared to 62-73p being reallocated in the treatment conditions.** This amounts to people deducting 27-36% as much revenue in the proxy condition as they would do were the bank avoiding tax. While people react to the use of information which could act as a proxy for other characteristics, this is equivalent to about a third of their aversion to tax avoidance - considerable but not overwhelming dissatisfaction.

In relation to perceived accuracy of the algorithms, we see that people who believe that Bank B will be more accurate are less likely to reallocate money away from them.

**Interestingly, when we prompt people to assume that Bank B is more accurate, we see a divergence in response (see graph below). Those that already believed that Bank B was more accurate reallocate more money to Bank B, particularly in the gender and ethnicity conditions. In contrast, people who did not think that Bank B was more accurate, do not change their allocation if they are in the gender or ethnicity prompt.**



**Finally, we see that the framing and ordering of information influences perceptions of fairness.** Firstly, when people are told that the algorithm is more accurate, they perceive its use more favourably. In theory, financial institutions could say that their algorithm is more accurate, with little evidence of accuracy, or overclaim its accuracy (particularly where such practices are not currently audited). Conversely, this information may be valuable for consumers in supporting the use of sophisticated technologies which act to increase equality. Secondly, we see that people exposed to the treatment conditions (i.e., the 'proxy' banks) punish the bank that avoids paying tax less than those in the neutral control condition, despite the fact that there is no difference between the questions that they were asked. It may be that when contentious practices by banks are considered cumulatively, people assume they are the norm and adjust their levels of acceptability.



## Recommendations

### Recommendations for the CDEI

1. Consider further testing to understand the impact of framing on perceptions of fairness, acceptance and comprehension of algorithm use, e.g:
  - Vary the framing of the use of information which may act as a proxy;
  - Vary how algorithm accuracy is framed;
  - Consider their interaction with each other.
2. On the use of social media data in bank algorithmic decision-making:
  - Collate a clearer understanding of its current or potential use;
  - Provide consumers with this information and further test how they perceive its use;
  - Ensure policy either protects or empowers consumers to get the most appropriate outcome (relevant to broader policy recommendations).

### Recommendations for policy makers and financial services

3. Encourage financial institutions who use algorithms to screen customers to test whether their systems may be (inadvertently) biased on the basis of gender or ethnicity, as a result of other information that acts as a proxy for the characteristics.
4. Consider whether policies to improve transparency as described in our above recommendations will, on their own, be sufficient, given that those most likely to be negatively affected by this issue perceive it as most unfair.
5. All financial institutions should be using sophisticated algorithms to better serve their customers, while considering point 1 above.
6. Consider the use of deliberative forums for consumers to feedback directly on its use on a regular basis.



# Part 1: Background and trial design

## 1.1 Background

Our partner for this project is the CDEI, which was created in late 2018 to advise government on how to leverage and react to data-driven technological change for the benefit of society. As part of this mandate, the CDEI is interested in building the (UK) **evidence base on how people perceive the fairness of online practices**, one area of which is the financial services sector.

This project aims to help build this evidence base, focusing on better understanding people's perceptions of fairness in online loan applications processes. More specifically, **the project will use a “spectator” economic game<sup>3</sup> to measure the perceived fairness of using algorithms to determine loan application outcomes**. Regulators such as the Financial Conduct Authority have increasingly engaged with fairness as a framework to guide their work.<sup>4 5</sup> Consumers are also concerned about their data being used to make financial decisions. In a 2018 US survey, the aggregation of personal finance scores using many types of consumer data was rated as unacceptable by 68% of respondents.<sup>6</sup> Further, financial services are often characterised by complex, difficult to compare products,<sup>7</sup> hence consumers' fairness perceptions may be all the more shrouded in uncertainty and biases.

The project will be conducted by the CDEI and by BIT's Predictiv and consumer policy teams, and will consist of an online experiment. The online experiment will use a “spectator” design to elicit participants' perceptions of the fairness of a bank's use of socio-economic data in algorithms.

<sup>3</sup> Our experimental design is detailed in this document and is adapted from: Cappelen, A. W., Moene, K. O., Sørensen, E. Ø., & Tungodden, B. (2013). Needs versus entitlements - an international fairness experiment. *Journal of the European Economic Association*, 11(3), 574-598; Cappelen, A. W., Luttens, R. I., Sørensen, E. Ø., & Tungodden, B. (2018). Fairness in Bankruptcies: An Experimental Study. *Management Science*; Almås, I., Cappelen, A., & Tungodden, B. (2019). Cutthroat capitalism versus cuddly socialism: Are Americans more meritocratic and efficiency-seeking than Scandinavians?. NHH Dept. of Economics Discussion Paper, (4).

<sup>4</sup> See Financial Conduct Authority. (2018). Fair pricing in financial services. *Discussion Paper*. DP18/9; and Competition & Markets Authority. (2018). Pricing algorithms. *Working Paper*. CMA94. (p.50).

<sup>5</sup> Financial Conduct Authority. (2018). Price discrimination in financial services: How should we deal with questions of fairness?. Research Note. Available at:

[https://www.fca.org.uk/publication/research/price\\_discrimination\\_in\\_financial\\_services.pdf](https://www.fca.org.uk/publication/research/price_discrimination_in_financial_services.pdf)

<sup>6</sup> In this survey, participants were presented with information on companies' ability to compute a “nontraditional credit score” from thousands of data points on consumers' activities and behaviours on the basis that “all data is credit data”. Pew Research Centre. (2018). Public Attitudes Toward Computer Algorithms. Report. Available at: <http://www.pewinternet.org/2018/11/16/public-attitudes-toward-computer-algorithms/>

<sup>7</sup> See Carlin, B. I. (2009). Strategic price complexity in retail financial markets. *Journal of Financial Economics*, 91(3), 278-287; and Shu, S. B. & Morelli, S. (2012). Applying fairness theories to financial decision-making. *Working paper*. Available at:

[http://www.anderson.ucla.edu/faculty\\_pages/suzanne.shu/Shu%20Morelli%20applying%20fairness.pdf](http://www.anderson.ucla.edu/faculty_pages/suzanne.shu/Shu%20Morelli%20applying%20fairness.pdf)

## 1.2 Project aims

**Social impact:** The social impact of this project will lie in its ability to define and quantify fairness perceptions in this context, so as to inform policy and further research. This project therefore has a diagnostic aim rather than a remedial one, as the experiment will seek to identify whether and how much there is an issue with how the fairness of online loan application processes is perceived by consumers, in particular the use of algorithms to assess applications. The results of this project will inform both the CDEI's recommendations to government on how to tackle such business practices, and further work on how to remedy fairness issues in this context.

**Overall aims:** There is currently a lack of evidence and understanding on how UK consumers think about online loan applications processes. This includes 1) whether they understand what kind of information is used to make decisions, and 2) how fair they think it is to use algorithms to assess applications. The project aims to help build a better understanding of these issues to inform policy advice. Its results will be used by CDEI to engage with financial institutions about their loan review processes.

**Purpose of using this experimental design:** An online experiment is an especially appropriate methodology for this research topic because we are seeking to better understand people's perceptions of an online process. Additionally, we will examine the impact of participants' demographic characteristics on their fairness perceptions, and using the Predictiv platform will help recruit a diverse group of participants in a short timeframe.

**Predictiv methodological checklist:** This experiment is set up as a diagnostic test. For Predictiv simulations, we normally check the external validity of the study by completing the [methodological checklist](#). Since this experiment does not use a decision simulation and instead focuses on an outcome not directly observable in the real world (i.e., fairness perceptions), we do not include this checklist in this Trial Report.

## 1.3 Online experiments and Predictiv

Predictiv ([www.predictiv.co.uk](http://www.predictiv.co.uk)) is an online platform for running behavioural experiments built by the Behavioural Insights Team. It enables governments and other organisations to run randomised controlled trials (RCTs) with an online population of participants, and to test whether new policies and interventions work before they are deployed in the real world.

Predictiv provides access to a large international panel, including more than 200,000 individuals in the UK and 2,000,000 in the US, as well as the functionality to run a range of online experiments.

More information on the methodology behind Predictiv, including payments, randomisation, recruitment, data storage and ethics can be found [here](#).

This trial follows these standard procedures with one exception. Individuals in the banker role will have to leave us their email address (personally identifiable information) in order to participate in the experiment. This is necessary to administer the variable payment, which is based on the spectator's decision, and which they receive after both roles (bankers and spectators) have participated. In addition, spectators have the option to leave us their email address so that we can send them a screenshot with the payment confirmation. This is to increase our credibility that spectator decisions are indeed consequential. Participants in the experiment will be clearly informed about the purposes of collecting personally identifiable information (PII) and be asked for their consent. Data handling and retention follows BIT protocols and the General Data Protection Regulation (GDPR).

## 1.4 Experimental design and procedures

In this experiment our objective is to use a robust measure of fairness perceptions and to test the effect of different prompts about informational proxies that a bank might be using on such fairness perceptions. To do so, we are adopting a variation of an established fairness experiment pioneered by Bertil Tungodden, Alexander Cappelen and co-authors.<sup>8</sup>

The experiment is a strategic game where a participant is asked to play the role of spectator, who observes a choice by two other players. After observing these decisions, the spectator has an opportunity to redistribute earnings between the two players. Note that this decision is consequential: it determines actual financial outcomes for the two individuals. We assess how much money the spectator is willing to reallocate to punish someone for (perceived) unfair behaviour, which we use as our proxy for fairness perceptions and compare this across treatments (see 'Analytical framework' below).

Below we describe the different stages of the experiment. In the following section we discuss the practical implementation of the design.

### Stage 1: Player assignment

There are three players in the game: two bankers and a spectator.  
All players are real respondents in the experiment.

All players receive general instructions about their roles and what they are asked to do in the experiment.

<sup>8</sup> Cappelen, A. W., Moene, K. O., Sørensen, E. Ø., & Tungodden, B. (2013). Needs versus entitlements—an international fairness experiment. *Journal of the European Economic Association*, 11(3), 574-598; Cappelen, A. W., Luttens, R. I., Sørensen, E. Ø., & Tungodden, B. (2018). Fairness in Bankruptcies: An Experimental Study. *Management Science*; Almås, I., Cappelen, A., & Tungodden, B. (2019). Cutthroat capitalism versus cuddly socialism: Are Americans more meritocratic and efficiency-seeking than Scandinavians?. *NHH Dept. of Economics Discussion Paper*, (4).


**Banker 1****Banker 2****Spectator****Stage 2: Choice by bankers**

The bankers are informed that they are senior executives and get to choose the bank's strategy and how it operates. They have two choices. One option (A) is a neutrally framed operating model of a standard bank. The other option (B) is a strategy where the bank uses predictive analytics to determine eligibility for credit, which are neutral or based on gender, ethnicity, or social media presence (this will depend on the treatment - see below).

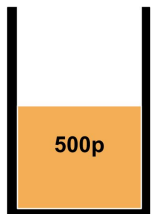
Choosing option A gives the banker £5; choosing option B gives the banker £6.50. The bankers are asked to make a total of four decisions, one for each of the four predictive analytics scenarios (which will be shown in a random order).

**Bank A****Bank B****Stage 3: Choice by spectator**

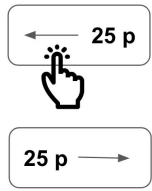
The spectator is informed about the choice of each of the bankers and their associated earnings. The spectator can then decide to reallocate earnings between the players (at 25p increments). Any reallocation is acceptable and the spectator does not incur a cost when reallocating earnings. The spectator can only reallocate funds - earnings cannot be destroyed or given back to the experimenter.

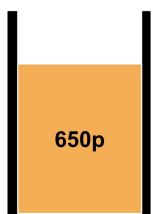


**Spectator**



**Banker 1: chose Bank A**





**Banker 2: chose Bank B**

**Stage 4: Additional questions**

1. Free-text question where spectators can explain why they made the allocation decision they did. The question reads: *It's very useful for us to understand the reasons why you chose this allocation. Please use the box below to give some details.*
2. Please think back to the scenario with the two bank executives. We are going to present you with another allocation choice. The scenario is hypothetical, so it won't affect the payment the bankers receive.
  - Imagine that Bank B is registered in the Cayman Islands and doesn't pay any corporation tax to the UK government.
  - If you had the chance to reallocate earnings between someone who chose Bank A and someone that chose Bank B, what would you choose?
3. Think back to the first allocation decision you made between Bank A and Bank B.
  - Imagine a situation where one bank uses advanced computing techniques to decide who receives a loan. Another bank does not use these techniques and instead uses a team of trained experts. How accurate do you think the assessments of the bank using the computing techniques will be compared to the bank with the trained experts?
4. Now suppose Bank B's computer techniques allow them to make more accurate predictions about applicants' reliability than Bank A. Would knowing this make you change your allocation choice between Bank A and Bank B? ((Show previous allocation and allow them to adjust))

\*\*\*

In this stage, respondents will also be asked additional demographic questions. These are listed in the covariates table.

**Stage 5: Earnings**

At the end of the experiment the bankers are paid according to the final redistribution decision of the spectator. Note that this makes their choice of bank strategic: the decision to choose Bank B over Bank A depends on how much they expect the spectator to care and punish them for this decision. The spectator receives a flat fee for participating. In other words, their redistribution decision does not impact her/his earnings from the experiment.

### 1.4.1 Practical implementation

We are employing various techniques to make the implementation of this experiment practically feasible. In particular, we need a workaround for the constraint that Predictiv does not yet have the functionality to connect respondents to each other in real time. However, even if this functionality were available<sup>9</sup>, simultaneous participation is difficult to implement online and suffers from inefficiencies. In particular, it is more likely that respondents drop out of the experiment mid-way when participating online compared to in a physical laboratory environment (for Predictiv experiments we assume a drop-out rate of roughly 15%). If a drop-out does happen, the experimenter needs to have a plan to 1) handle the session for the matched group (e.g., terminate the session and pay everyone; continue the session and replace the drop-out with a computer); and 2) handle the data of the matched group in the analysis (e.g., exclude the respondents in the matched group with the drop-out).

Fortunately, the use of the strategy method does not seem to significantly affect decisions<sup>10</sup>, nor do individuals change their decisions if only one or all decisions are paid out<sup>11</sup>. This makes it easier to decide in favour of the strategy method and avoid the complications listed above.

Another advantage of the strategy method is that it allows us to select cases for the spectator that are of interest for our study. In particular, this is the case where one banker chooses Bank A and the other chooses Bank B, rather than both choosing Bank A or both choosing Bank B. This approach has been used in other economic experiments for efficiency reasons<sup>12</sup>, including work using the spectator design<sup>13</sup>.

The session order of the experiment sessions will be as follows:

- **Session 1 - bankers:** 760 individuals will be recruited to play the role of bank executive. They are given general instructions, which includes information that their choice will be observed by a spectator who can then decide to reallocate earnings between them and another executive, who also made a decision and received the same instructions. In order to be eligible for the reward, individuals need to give us their email address so that we can send them their reward voucher. We will elicit email addresses at the beginning of the experiment so that we do not have any

<sup>9</sup> There are some options available, such as: Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics*, 21(1), 99-131.

<sup>10</sup> Brandts, J., & Charness, G. (2000). Hot vs. cold: Sequential responses and preference stability in experimental games. *Experimental Economics*, 2(3), 227-238; Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, 14(3), 375-398.

<sup>11</sup> Charness, G., Gneezy, U., & Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, 131, 141-150.

<sup>12</sup> For example: Gülerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312(5770); Gülerk, Ö., Irlenbusch, B., & Rockenbach, B. (2014). On cooperation in open communities. *Journal of Public Economics*, 120, 220-230.

<sup>13</sup> Cappelen, A. W., Falch, R., & Tungodden, B. (2019). The Boy Crisis: Experimental Evidence on the Acceptance of Males Falling Behind. *NHH Dept. of Economics Discussion Paper*, (06).

missing contact information for any individuals in the study. This means that if respondents do not want to give us their email address, they will not be allowed to continue. For efficiency reasons, each respondent will go through each of the 4 different treatment scenarios (neutral phrasing, selecting on gender, selecting on ethnicity, selecting on social media presence) in random order, and will be asked to make a decision for each. This will facilitate the matching in the next stage. Respondents are informed that one of the scenarios will be selected for payment and that they will receive payment depending on their choice and that of the spectator in any case in the chosen scenario.

- Matching: We are looking to create A+B choice pairings in each of the scenarios. We will match bankers to pairs for each condition. This means that we will have 95 unique pairs per condition ( $(720 / 2) / 4$ ).
- Session 2 - spectators: 1800 spectators (450 per condition) are recruited for the experiment. They are told that they have been randomly matched to a pair of bank executives that completed their choice experiment earlier this week. After the instructions, spectators will be informed of the bankers' choices and can make their reallocation decision. They are informed that 1 out of 5 spectators will be randomly selected to have their choice implemented and paid. They can choose to leave their email address if they want to receive a screenshot of the payment to the players in the banker role.
- Payment calculation and transfer: After the experiment, payments for each banker are calculated using Stata/R. Amazon vouchers with the correct denomination will be generated using Predictiv's payment systems and recipients will receive their code via email. Where applicable, corresponding spectators will receive a screenshot of the payment. Email addresses of respondents in the experiment will be deleted immediately afterwards.

In Predictiv experiments, we provide truthful instructions to participants. Since we want to communicate to spectators that they are matched to a unique pair of bankers, it is important that we have a successful matching algorithm. We will use the following procedure:

- Once the bankers session is completed, we will begin the matching procedure for the condition that has the lowest proportion of A choices.
- For this condition, we will match all A bankers to a participant that chose bank B in that condition.
  - Note: given that we need a minimum of 90 A+B pairs per condition, we need at least 12.5% of the bankers to choose A in the lowest proportion condition. If this threshold is not met, we will recruit additional players into the banker role. In the pilot experiment, we will also assess the impact of reducing the monetary benefit of choosing Bank B.
- Once the pairs for the condition with the lowest proportion of As have been made, we will move on to the condition that has the second lowest proportion of A choices. The same matching procedure is applied for this condition. At this stage, we will give priority to matching A bankers that have also chosen A in any of the other remaining



scenarios. This lowers the probability that we end up with excess bankers (e.g., only B+B pairs).

- When we have gone through all conditions, any remaining bankers (most likely B+B pairs) will be matched. These pairs will be presented to spectators in a separate session (Session 2b - spectators), where the spectators will be presented with a B+B banker pair instead. This allows us to match all excess bankers and keep our instructions truthful (and simple) to spectators.

In case of any technical difficulties with the implementation (e.g., over-recruitment of spectators), we will randomly select an A+B banker pair and pay this pair out for the other spectator's decision as well. This means that these bankers receive payment twice (one for each spectator decision).

### 1.4.2 Behavioural predictions

Following the model in the original spectator game<sup>14</sup>, we assume that the utility function of the spectator for banker  $j$  contains two main elements: 1) an income inequality parameter  $\lambda$  that is negative and linearly increasing by the level of the income difference between bankers  $j$  and  $k$ ; 2) a fairness parameter  $\gamma$  that is negative and non-zero when banker  $j$  chooses Bank B. Note that  $\gamma$  is a function of the treatment condition, as well as whether banker  $k$  chooses Bank A when banker  $j$  chooses Bank B.

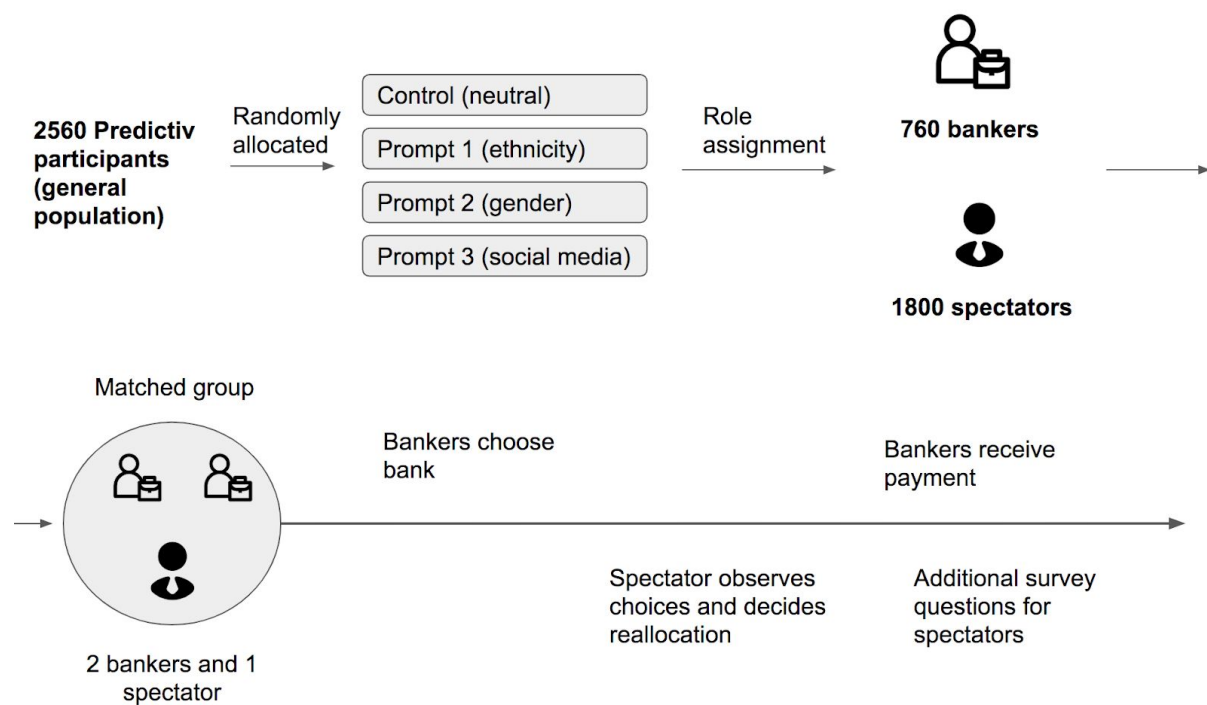
This utility function yields the following predictions:

- In the control/neutral condition, we assume fairness concerns play a limited role. Nonetheless, it could be relevant for some participants, for example because of privacy concerns and/or suspicion about general use of technology in combination with personal data. In addition, the spectator may redistribute earnings if they are inequality averse, but we expect this to be minor.
- In the treatment conditions, we expect the spectator to redistribute more money away from the banker choosing Bank B towards the person choosing Bank A if the prompt has moral weight. Depending on the weight of the spectator's fairness parameter, the spectator may allocate more than equal points to the banker choosing Bank A to punish the banker choosing Bank B.<sup>15</sup>

---

<sup>14</sup> Cappelen, A. W., Moene, K. O., Sørensen, E. Ø., & Tungodden, B. (2013). Needs versus entitlements—an international fairness experiment. *Journal of the European Economic Association*, 11(3), 574-598.

<sup>15</sup> The literature on public good dilemmas finds that individuals are indeed willing to financially punish others for fairness motivations, even if they do not derive any (future) benefit from doing so and even if this mechanism is costly for them to apply. The seminal paper on this is Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4), 980-994.



**Figure 1: Test overview**

### 1.4.3 Pre-test

We are not conducting a pre-test for this project. However, we will run a pilot with the bankers to get a rough steer on what proportion are choosing Bank A in each condition. If this is very skewed (e.g., 10% choose Bank A; 90% choose Bank B), then we may need to recruit more individuals into the banker role in order to end up with enough respondents to match with the spectators. We expect the control condition to be the most skewed: since the prompt for Bank B is neutral, we do not expect many bankers to choose Bank A and forgo higher payment.

## 1.5 Treatments and randomisation

In the experiment, we will be varying the information associated with Bank B. This is detailed in the table below. By contrast, the prompt for Bank A is neutral and reads: *Bank A uses financial information to determine an individual's application, such as a person's salary, whether someone is employed and if they have debt.*

**Table 1: Overview of the different conditions in the test**

Condition	Description	N
Control (neutral)	Bank B uses advanced computing techniques and a broader range of personal information than Bank A to make decisions about loan applications. This allows it to make predictions about an individual's application.	450

Gender	Bank B uses advanced computing techniques and a broader range of personal information than Bank A to make decisions about loan applications. This allows it to make predictions about an individual's application. However, by including information which may link to gender, for example, salary and occupation, Bank B may end up offering different levels of credit to men and women.	450
Ethnicity	Bank B uses advanced computing techniques and a broader range of personal information than Bank A to make decisions about loan applications. This allows it to make predictions about an individual's application. However, by including information which may link to ethnicity, for example, salary, postcode and occupation, Bank B may end up offering different levels of credit to people with different ethnicities.	450
Social media	Bank B uses advanced computing techniques and a broader range of personal information to make decisions about loan applications than Bank A. This can include information from applicants' social media profiles and it allows Bank B to make predictions about applicants' reliability. However, by including social media data, Bank B may end up offering different levels of credit to people with a greater social media presence compared to people with a smaller social media presence.	450
<b>TOTAL</b>		<b>1800</b>

Participants are randomly assigned to a treatment at an individual level. When a participant enters the experiment, they are given a random number representing an intervention. Depending on the number assigned, they are taken through a separate path in the experiment that corresponds with a specific intervention (e.g., prompt that ethnicity information is used to review an application). The random number is stored in the data output and used for data analysis to assess the intervention's impact on the outcome variables.

The full experimental instructions and treatment material can be found on the Predictiv dashboard.

## 1.6 Participant pool and eligibility

All respondents for the study are drawn from the general population, which means that all participants in the UK panel are eligible.

Using the platform's security checks, we will ensure that participants are unique across experiments (i.e., it will not be possible for a banker to also participate in the experiment as a spectator). To optimise the viewing experience of the spectators, we will exclude participants that are accessing the study on a mobile device.

## 1.7 Outcome measures

The table below lists the outcome measures and covariates used in this trial. The second column explains how each variable is constructed, the third column details the coding (if applicable) and the final column indicates whether it will be used for primary or secondary analysis.

**Table 2: Outcome measures**

PRIMARY		
Measure	Definition	Coding
<i>Amount reallocated</i>	<i>The amount in pence that the spectator reallocates from the banker choosing Bank B to the banker choosing Bank A.</i>	<i>Continuous variable that is bounded between -500 and +650.</i>
EXPLORATORY		
<i>Fairness benchmark (for descriptive analysis)</i>	<i>The hypothetical amount reallocated from the banker choosing Bank B to banker choosing Bank A in the unethical practice scenario. This helps us interpret the magnitude of treatment effects on our primary outcome measure.</i>	<i>Continuous variable that is bounded between -500 and +650.</i>
<i>Motivations</i>	<i>It's very useful for us to understand the reasons why you chose this allocation. Please use the box below to give some details.</i>	<i>Free-text answer.</i>

<i>Perceived accuracy 1</i>	<i>Imagine a situation where Bank A uses advanced computing techniques to decide who receives a loan. Another bank, Bank B, uses less advanced computing techniques and draws on a smaller pool of data. Both banks are trying to make decisions about whether to give someone a loan or not. How accurate do you think the assessments of the bank using the advanced computing techniques (Bank A) will be compared to Bank B?"</i>	<i>Scale variable ranging from -5 to 5: -5 → Bank A will be much more accurate than Bank B; 0 → The banks will have the same accuracy; 5 → Bank B will be much more accurate than Bank A.</i>
<i>Perceived accuracy 2</i>	<i>The amount of money (p) taken from Bank B and given to Bank A by the spectator when explicitly informed that advanced computing techniques increase Bank B's predictions about applicants' reliability.</i>	<i>Continuous variable that is bounded between -500 and +650.</i>

In addition, we collect the following covariates. Note that the vector column indicates which vector these variables belong to in the regression analysis (see 'Analysis strategy').

**Table 3: Description of covariates**

<b>COVARIATES</b>			
<b>Measure</b>	<b>Vector</b>	<b>Definition</b>	<b>Coding</b>
<i>Treatment</i>		<i>Treatment assignment</i>	<i>Categorical variable: Control → 0 Gender → 1 Ethnicity → 2 Social media → 3</i>
<i>Gender</i>	<i>A</i>	<i>"What is your gender?" *</i>	<i>Categorical variable: Male → 0 Female → 1</i>
<i>Age</i>	<i>A</i>	<i>"What is your age?" *</i>	<i>Categorical variable: 18-24 → 1 25-39 → 2 40-54 → 3 55+ → 4</i>

<i>Household income</i>	A	<i>“What is your current annual household income before taxes?” *</i>	<i>Categorical variable based on median income in UK: &lt; £27,999 → 1 ≥ 28,000 → 2</i>
<i>Location</i>	A	<i>“In which region do you live?” * ; Original variable has 12 levels. (NUTS1).</i>	<i>Categorical variable: London → London, 0 North East; North West; Yorkshire &amp; Humber → North, 1 East of England; South East; South West → South &amp; East, 2 East Midlands; West Midlands → Midlands, 3 Wales, Scotland, N. Ireland → Wales, Scotland &amp; N. Ireland, 4</i>
<i>Education level</i>	A	<i>“What is the highest education level that you have achieved?”</i>	<i>Categorical variable: None → 0 Secondary school → 1 Post-secondary → 2 Vocational → 3 Undergraduate → 4 Prof. qualification → 5 Postgraduate → 6</i>
<i>Ethnicity</i>	A	<i>“What is your ethnic group? Choose one option that best describes your ethnic group or background.”</i>	<i>Categorical variable: White → 0 Mixed/multiple ethnic groups → 1 Asian / Asian British → 2 Black/African/Caribbean/Black British → 3 Other ethnic group → 4</i>
<i>Experience - loan 1</i>	E	<i>“Have you ever applied for a bank loan?”</i>	<i>No → 0 Yes → 1 Don't know → 2 Prefer not to say → 3</i>
<i>Experience - loan 2</i>	E	<i>“What was the outcome of your application? In other words, did you receive the loan you requested?” [CONDITIONAL ON EXPERIENCE-LOAN 1 ANSWER BEING 'YES']</i>	<i>Yes → 0 Yes, with some changes → 1 No → 2 Don't know/remember → 3 Prefer not to say → 4</i>
<i>Experience - loan 3</i>	E	<i>“Since you have experience with loan applications, it is helpful for us to understand how this application process was for you. Use the box below to describe your</i>	<i>FREE TEXT ANSWER</i>

		<i>experience.” [FREE TEXT BOX] [CONDITIONAL ON EXPERIENCE-LOAN 1 ANSWER BEING ‘YES’]</i>	
<i>Social media use 1</i>	<i>R</i>	<i>“How many social media platforms are you active on (e.g., Facebook, Instagram, LinkedIn, Snapchat)?”</i>	<i>Ordinal variable: None → 0 1 → 1 2-3 → 2 4-6 → 3 7+ → 4</i>
<i>Social media use 2</i>	<i>R</i>	<i>“How often do you access social media?”</i>	<i>Ordinal variable: Never → 0 Once a week → 1 A couple of times a week → 2 Once a day → 3 A couple of times a day → 4 Many times a day → 5</i>
<i>Digital literacy</i>	<i>R</i>	<i>“On a scale of 1 to 5, how would you rate your computer skills?” [SCALE, where 1: I have difficulty understanding how to use computers and need assistance to use them”; 3 “I am confident using computers for some purposes but not others”; “5: I am very comfortable with computers and know how to use them for a wide variety of purposes”</i>	<i>Ordinal variable coded 1 through 5.</i>
<i>Payment credibility</i>	<i>R</i>	<i>“In this study we’ve told you that we will randomly select 1 banker pair out of every 5 pairs and pay them the earnings that their spectator chose for them. We are committed to this, but want to ask you if you believe that we will and can follow through with this payment? (If you don’t, then this is important for us to know for future studies). Please choose the option that is closest to what you believe.” 0 - I <u>don’t believe</u> that players in the banker role will be paid; 1 - I <u>do believe</u> that players in the banker role will be paid.</i>	<i>Categorical variable: Don’t believe → 0 Believe → 1</i>
* Participants are automatically profiled on standard demographic characteristics (age, gender, location, income), which means that this information does not need to be solicited in the experiment.			



## 1.8 Analysis strategy

### 1.8.1 Primary Analysis

The primary analysis focused on the amount of money (in pence) that is reallocated between the bankers by the spectator. By comparing this measure across treatments, we will determine which informational proxy is perceived as most unfair relative to the neutral baseline.

*Equation 1:*

$$Reallocated_i = \alpha + \beta_1 Treatment_i + A_i\Gamma + \varepsilon_i$$

where:

$Reallocated_i$  is a continuous variable (bounded at -500 and +650) that represents the amount in pence that is reallocated by spectator  $i$  from the banker choosing Bank B to the banker choosing Bank A.

$\alpha$  is the regression constant.

$Treatment_i$  is a vector of binary indicators with a value of 1 if participant  $i$  was assigned to the treatment and 0 otherwise.

$A_i$  is a vector of controls which indicate the gender, age bracket, income bracket, location, ethnicity and education of participant  $i$ . These variables are treated as dummy variables and are coded as stated in the covariates table.

$\varepsilon_i$  is the error term.

### 1.8.2 Secondary Analysis

In the secondary analysis, we test for specific subgroup effects. In particular, we assess whether women react more strongly than men to the informational proxy of gender compared to the control condition. For ethnicity, we assess whether non-white individuals react more strongly than white individuals to the informational proxy of ethnicity.

*Equation 2a:*

$$Reallocated_i = \alpha + \beta_1 T\_Gender_i + \beta_2 T\_Gender_i * Gender_i + A_i\Gamma + \varepsilon_i$$

*Equation 2b:*

$$Reallocated_i = \alpha + \beta_1 T\_Ethnicity_i + \beta_2 T\_Ethnicity_i * Nonwhite_i + A_i\Gamma + \varepsilon_i$$

Equation 2c:

$$Reallocated_i = \alpha + \beta_1 T\_SocialMedia_i + \beta_2 T\_SocialMedia * SocialMediaUse_i + \beta_3 SocialMediaUse_i + A_i \Gamma + \epsilon_i$$

Equation 2d:

$$Reallocated_i = \alpha + \beta_1 T\_SocialMedia_i + \beta_2 T\_SocialMedia * Age_i + A_i \Gamma + \epsilon_i$$

where:

$Reallocated_i$  is a continuous variable (bounded at -500 and +650) that represents the amount in pence that is reallocated by spectator  $i$  from the banker choosing Bank B to the banker choosing Bank A.

$\alpha$  is the regression constant.

$T\_Gender_i$  is a binary indicator with a value of 1 if participant  $i$  was assigned to the gender prompt treatment and 0 otherwise.

$T\_Gender * Gender_i$  is an interaction term between  $T\_Gender_i$  and  $Gender_i$ , which has a value of 0 if the participant is male and a value of 1 if the participant is female.

$T\_Ethnicity_i$  is a binary indicator with a value of 1 if participant  $i$  was assigned to the ethnicity prompt treatment and 0 otherwise.

$T\_Ethnicity * Nonwhite_i$  is an interaction term between  $T\_Ethnicity_i$  and  $Nonwhite_i$ , which has a value of 0 if the participant is white and a value of 1 if the participant is non-white.

$T\_SocialMedia_i$  is a binary indicator with a value of 1 if participant  $i$  was assigned to the social media prompt treatment and 0 otherwise.

$T\_SocialMedia * SocialMediaUse_i$  is an interaction term between  $T\_SocialMedia_i$  and  $SocialMediaUse_i$ , which is a categorical variable that indicates the frequency with which participant  $i$  uses social media.

$SocialMediaUse_i$  is a categorical variable that indicates the frequency with which participant  $i$  uses social media.

$T\_SocialMedia * Age_i$  is an interaction term between  $T\_SocialMedia_i$  and  $Age_i$ , which is a categorical variable indicating the age bracket of participant  $i$ .

$A_i$  is a vector of controls which indicate the gender, age bracket, income bracket, location, ethnicity and education of participant  $i$ . These variables are treated as dummy variables and are coded as stated in the covariates table.

$\varepsilon_i$  is the error term.

### 1.8.3 Robustness checks

We will run robustness checks by repeating our models from the primary and secondary analysis with additional covariates in vector R.

### 1.8.4 Exploratory Analysis

We will run exploratory analysis on the variables listed in the corresponding table in the ‘outcome measures’ section as well as the covariates in vector E.

In addition, we will rerun our main analysis as a log-linear model to assess whether this fits the data better than the OLS model. We will evaluate this by comparing the R-squared across the two models.

## 1.9 Power calculations

BIT runs power calculations for every trial to assess whether we can be sufficiently confident that we can detect a difference between the intervention and the control material. This is based on the number of individuals participating in each of the test conditions, the variance in responses, and insights from academic literature and previous studies on the impact of the intervention tested.

In our power calculations, we follow current best practice<sup>16</sup> by adopting a significance threshold for the p-value of our statistical tests of 5%. In addition, we aim to have sufficient statistical power to detect an effect, should it exist, with 80% confidence.

We run our calculations for a range of standard deviations and sample sizes that can be accommodated within the project budget. We include a range of standard deviations, though our focus is on the effect size measured as Cohen’s D.

At the budgeted sample of 1800 spectators, we are looking at an effect size of 0.187 Cohen’s D. We feel this is acceptable given the controlled environment.

**Table 4: Summary of Power Calculations**

Sample size	Number of arms	Baseline SD (pence)	Effect size (Cohen’s D)	Effect size (substantive)
1800	4	50	0.18697393	9.3486965
1800	4	100	0.18697393	18.697393

<sup>16</sup> List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14(4), 439.

1800	4	150	0.18697393	28.0460895
2000	4	50	0.1773605	8.868025
2000	4	100	0.1773605	17.73605
2000	4	150	0.1773605	26.604075
2200	4	50	0.16908924	8.45446177
2200	4	100	0.16908924	16.9089235
2200	4	150	0.16908924	25.3633853
2400	4	50	0.1618792	8.09396019
2400	4	100	0.1618792	16.1879204
2400	4	150	0.1618792	24.2818806
2600	4	50	0.15549753	7.77487651
2600	4	100	0.15549753	15.549753
2600	4	150	0.15549753	23.3246295

## 1.10 Risks

Below we list the main risks for the trial and our strategy to mitigate them.

**Table 5: Key risks**

Risk	Strategy to mitigate risk	Responsibility	Timeframe (if applicable)
Respondents interact with the material in such a way such that we see no variation across any of the prompt conditions (ceiling/floor effects). In particular, 1) it could be that spectators always choose to redistribute all the money to Bank A or do not redistribute any money; 2) all bankers choose only Bank A or Bank B.	We are running a pilot to assess whether floor/ceiling effects are occurring in the final design. In addition, we will be able to determine if this is due to a wider issue with the experimental design or our treatments by looking at the benchmarking questions.	BIT	Before trial launch
Spectators may not believe that their decisions are actually consequential for the payment of the bankers.	We will use clear instructions to emphasise that the decisions of the spectator are consequential. In addition, spectators will be given the opportunity to receive a screenshot with the payment transfer to their email. This should add credibility to our claim that we will pay out based on their allocation decision.	BIT	Trial design stage

## Part 2: Findings and recommendations

---

### 2.1 Implementation

The banker session ran between 19 August 2019 and 22 August 2019. The spectator session ran between 23 August 2019 and 9 September 2019 (14 days) and we collected data from 1828 spectators. We capped entry into our treatments towards the end of the data collection period such that we would avoid oversampling individuals beyond the required 450 per treatment. Since spectators are matched to a limited number of bankers, it would cause problems for our matching if we oversampled. The trial was implemented as described in the above Trial Protocol.

In total, we obtained the required 360 A-B banker pairs (exactly 90 per treatment) to match to 1800 spectators. We then had 20 excess pairs, 7 of which were also A-B pairs and 13 of these were A-A pairs. The matching procedure was executed as described in this trial protocol (the code is included in Appendix D). The A-A banker pairs were matched to 7 spectators in a separate session which does not contribute data to this analysis.

### 2.2 Results

#### 2.2.1 Differential attrition and balance checks

In both sessions we saw an above-average level of attrition compared with other Predictiv trials, which are predominantly happening on the landing page where respondents are asked for their email address. For example, in the spectator session, the attrition rate is 32.7%. Out of these, 60% drop out on the opening screen. We hypothesise that this is related to people not wanting to give us their email address to take part in the experiment, which is a fair concern. In any case, the bulk of the attrition happens before being exposed to our treatment variation. Attrition, both conditional on treatment exposure and not, does not differ significantly across treatments (the lowest p-value in these pairwise comparisons is  $p=0.312$ ).

Participants were balanced across treatments and control in terms of gender ( $\chi^2 = 2.55$ ,  $p = 0.47$ ), age ( $\chi^2 = 8.66$ ,  $p = 0.47$ ), household income ( $\chi^2 = 0.50$ ,  $p = 0.92$ ) and location ( $\chi^2 = 39.38$ ,  $p = 0.21$ ).

## 2.2.2. Descriptive statistics

**Table 6: Sample breakdown on key demographics**

	% of sample
<b>Gender</b>	
Female	48.6
Male	51.4
<b>Age</b>	
18-24 years	27.6
25-39 years	25.1
40-54 years	21.2
55 and over	26.1
<b>Income</b>	
Below median ( $\leq$ £27,499)	50.1
Above median ( $\geq$ £27,500)	49.9
<b>Location</b>	
London	15.3
North	23.4
South & East	31.0
Midlands	15.9
Wales, Scotland & N Ireland	14.4
<b>Education (highest)</b>	
None	1.4
GSCE	20.8
A-level	20.2
Vocational diploma	15.9
Undergraduate	26.7
Professional qualification	4.8
Postgraduate	10.1
<b>Ethnicity</b>	
White	87.1
Mixed/multiple ethnic groups	2.5
Asian/Asian British	6.1
Black/African/Caribbean/Black British	3.7
Other	0.7

**Table 7: Sample breakdown on social media use and previous loan experience**

	% of sample
<b>Social media use - platform #</b>	
None	9.2
1	21.4
2-3	44.1
4-6	22.8
7 or more	2.5
<b>Social media use - frequency</b>	
Never	8.6
Once a week	9.1
Couple of times a week	9.4
Once a day	11.9
Couple of times a day	21.8
Many times a day	39.2
<b>Computer skills</b>	
1 - difficulty understanding	0.9
2	2.4
3 - comfortable	25.6
4	25.4
5 - very comfortable	45.7
<b>Previous loan experience</b>	
Has applied for a loan	47.3
Has not applied for a loan	49.3
Don't know/prefer not to say	3.3

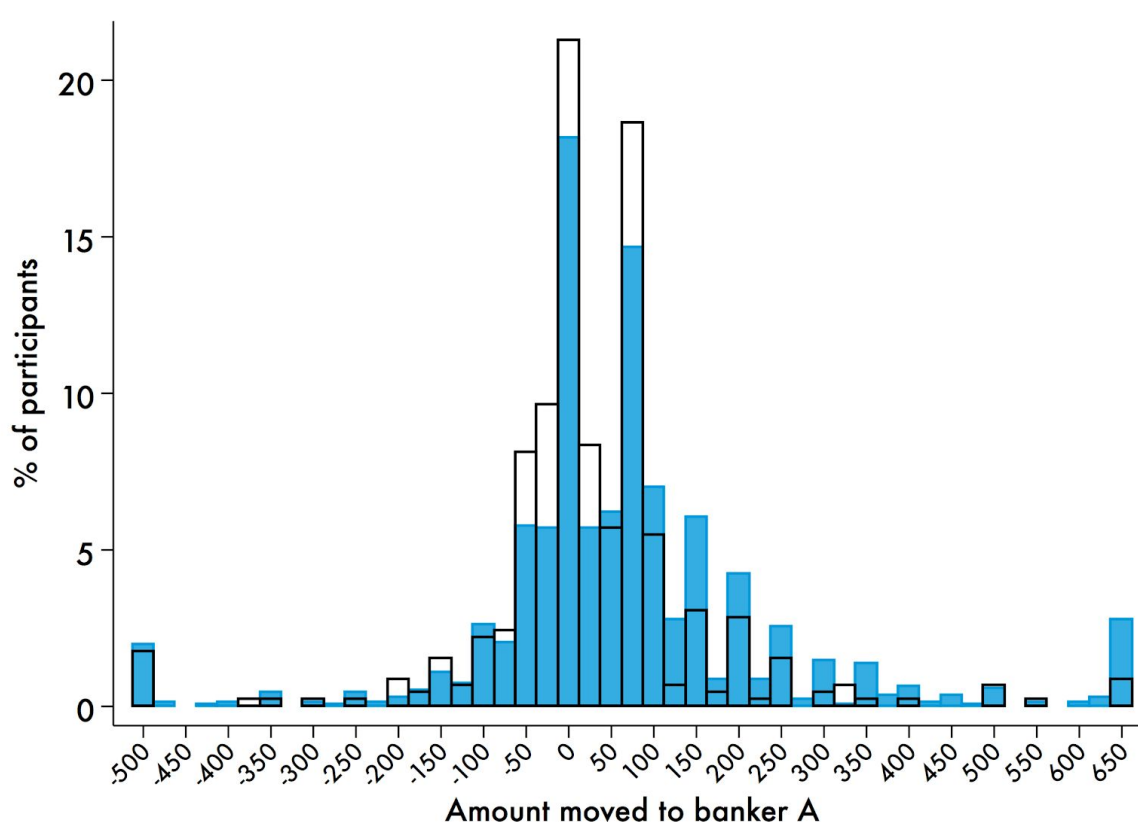
**Table 8: Descriptive statistics of primary and secondary outcome variables (standard deviation in brackets)**

	Neutral	Gender	Ethnicity	Social media
<b>Amount reallocated (pence)</b>	28.23 (136.92)	62.42 (176.22)	62.83 (179.99)	75.54 (181.93)
<b>Benchmarking</b>				
Amount reallocated (pence; tax avoidance scenario)	155.26 (258.02)	130.66 (268.16)	141.19 (251.00)	107.73 (272.29)
Amount reallocated (pence; assuming accuracy)	-3.84 (177.30)	44.39 (196.50)	46.14 (193.95)	49.18 (200.42)
Perceived accuracy (-5 to 5 scale)	1.70 (2.81)	1.48 (2.79)	1.52 (2.90)	0.84 (3.20)
N	456	459	454	459



### 2.2.3 Primary analysis

Figure 2 shows the distribution of reallocated money (in pence) from the banker that chose Bank B to the banker that chose Bank A (recall: the starting allocation is 500 pence for banker A and 650 pence for banker B). The control condition is shown by the bars with the black outline. The blue bars show the aggregate distribution of the three treatments. As will be discussed below, we find very little differences between each of our treatments. Overall, we see a shift in the distribution of reallocation to the right, meaning that a larger proportion of spectators are moving money towards banker A.



**Figure 2: The distribution of the amount reallocated from banker B to banker A in the control condition (black outline) and the treatment (blue)**

Table 9, which summarises the proportion of different reallocation ‘types’ shows this result in more detail, and by treatment. In the neutral prompt, 21.3% of spectators take no action (reallocating 0, thus keeping the starting allocation), 50.2% move some money towards banker A and the remaining 28.5% move money towards B. In the treatments we see movement across all of these categories. Specifically, there is a small decrease in the proportion of spectators taking no action (to 17-18%) and in the proportion that moves money towards banker B (to 21-23%). There is an increase in the proportion of spectators

allocating money to banker A (to 59-62%). Most of this increase in money towards banker A comes from spectators that are reallocating more than 75p, meaning that banker A receives more than half the available allocation.

The two spikes in the distribution (at 0 and at 75p), reflect two particularly salient decision options. A reallocation of 0 corresponds to the default allocation. Reallocating 75p to banker A creates pay equality between the two bankers, which is a common preference for individuals when redistributing income<sup>17</sup>.

**Table 9: Classification of types across treatments**

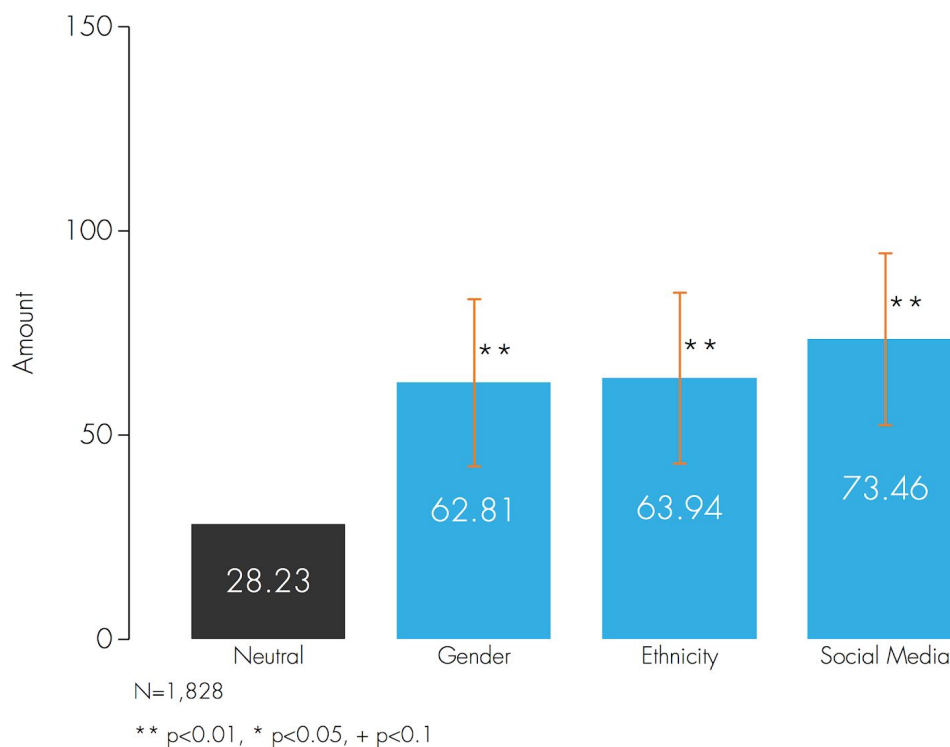
	No action	Money towards A				Money towards B
		Overall	Less than 75p	Equalising pay (75p)	More than 75p	
<b>Control</b>	<b>21.3%</b> <b>(N=97)</b>	<b>50.2%</b> <b>(N=229)</b>	<b>14.0%</b> <b>(N=64)</b>	<b>18.6%</b> <b>(N=85)</b>	<b>17.5%</b> <b>(N=80)</b>	<b>28.5%</b> <b>(N=130)</b>
<b>Treatments (overall)</b>	<b>18.1%</b> <b>(N=249)</b>	<b>59.5%</b> <b>(N=817)</b>	<b>11.9%</b> <b>(N=163)</b>	<b>14.7%</b> <b>(N=201)</b>	<b>33.0%</b> <b>(N=453)</b>	<b>22.3%</b> <b>(N=306)</b>
Gender	17.4% (N=80)	61.9% (N=281)	13.3% (N=61)	16.1% (N=74)	31.8% (N=146)	20.7% (N=98)
Ethnicity	18.3% (N=83)	58.6% (N=266)	10.8% (N=49)	14.3% (N=65)	33.5% (N=152)	23.1% (N=105)
Social media	18.7% (N=86)	58.8% (N=270)	11.5% (N=53)	13.5% (N=62)	33.8% (N=155)	22.4% (N=103)

The results from regression analyses, reported in Table 10, show that the amount reallocated in the treatments is significantly higher than in the control (by 35 to 45 pence). We do not find statistically significant differences in the reallocated amount between the three different prompts. This suggests that the fairness of these different uses of personal information is weighted similarly by spectators, on average. Interestingly, we do not find that any of our demographic characteristics (age, gender, household income, ethnicity, education, location) correlate significantly with the amount that is moved to banker A.

<sup>17</sup> Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, 94(4), 857-869.

**Table 10: Primary regression results showing the effect of the algorithm prompt on the amount reallocated.**

	Model 1: Comparisons against Neutral	Model 2: Comparisons against Gender	Model 3: Comparisons against Ethnicity
	Amount reallocated (pence)	Amount reallocated (pence)	Amount reallocated (pence)
<b>Treatment</b> (baseline = Neutral)			
Gender prompt	34.571** (10.44)		-1.133 (11.87)
Ethnicity prompt	35.704** (10.67)	1.133 (11.87)	
Social Media prompt	45.226** (10.74)	10.654 (11.72)	9.522 (11.99)
Neutral prompt		-34.571** (10.44)	-35.704** (10.67)
Demographic controls	YES	YES	YES
Constant	30.843 (30.64)	65.414* (30.51)	66.547* (30.54)
Observations	1,828	1,828	1,828
R-squared	0.023	0.023	0.023
Robust standard errors in parentheses; ** p<0.01, * p<0.05, + p<0.1			



**Figure 3: Primary analysis - the effect of treatment on the amount reallocated to banker A (controlling for demographic characteristics of the spectator)**

## 2.2.4 Secondary analysis

The secondary analysis explores subgroup effects. In particular, we are interested in whether:

- Women respond more strongly than men to the gender prompt compared to the other prompt conditions;
- Individuals from ethnic minority respond more strongly than white individuals to the ethnicity prompt compared to the other prompt conditions;
- Individuals with a stronger social media profile respond more strongly than individuals who are less present on social media to the social media prompt compared to the other prompt conditions. We had specified the frequency of social media platform use, as well as their age bracket, as proxies for social media presence.

In the sections below, we present the results for each of these subgroups verbally based on regression analyses. The regression outputs can be found in the appendix.

### 2.2.4.1 Gender

We find no significant differences in reallocated amount by gender. Directionally, women seem to reallocate more than men in the gender treatment compared to non-gender treatments, but this effect is not statistically significant ( $p=0.165$ ). It is worth noting that this analysis has limited power, and such an effect would have to be at least as big as the main

treatment effect for us to be able to reliably detect it. As such, these findings should be interpreted with caution and, ideally, would be investigated in a new trial that is explicitly powered to pick up on interaction effects.

#### **2.2.4.2. Ethnicity**

We find no significant differences in reallocated amount by ethnicity. Directionally, individuals who are white seem to reallocate slightly less than non-white individuals overall, but the difference is minimal in the ethnicity treatment. Note that this analysis has very low power; we would only have reliably detected a difference in treatment effect if it were over 100p.

#### **2.2.4.3. Social media**

We find no significant differences in reallocated amount by frequency of social media use. For age bracket, we find a significant interaction term for the age bracket 40-54 and a weakly significant effect for the 25-39 age group. These age groups reallocate more in the social media prompt condition than in other prompt conditions compared to individuals in the 18-24 group. These results are puzzling. Firstly, it is not immediately obvious why there would be a significant interaction of age, but not of social media frequency, when we hypothesised that the strength of the reaction would depend on how active someone is on social media. Stated differently, the frequency with which someone uses social media should be a more precise measure of this than someone's age bracket. In addition, it is also not clear why the 25-54 group would have a stronger reaction to the use of social media data compared to the 18-24 group, when the latter are likely the most active on social media. This suggests that the relationship between age and response in the social media treatment is likely spurious. It is worth exploring these effects more systematically in future research.

#### **2.2.4.4. Robustness checks**

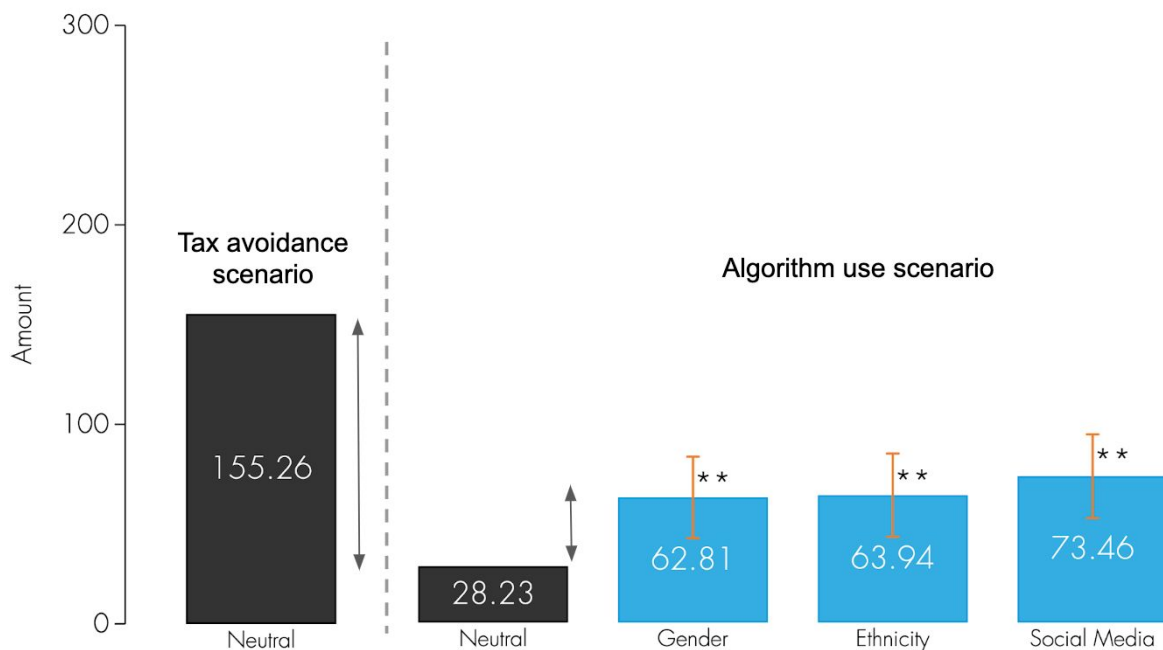
We reran our primary and secondary analyses with additional covariates on social media use and payment credibility (i.e., whether people believed that we would actually pay the bankers). We do not find significant correlations between the additional covariates and reallocation decisions with the exception of payment credibility. On average, people who do believe that we will pay the bankers (79% of the sample) reallocate 30p more than those that do not. Including this covariate in the regression does not change the results reported in the primary or secondary analyses.

### **2.2.5 Further exploratory analysis**

#### **2.2.5.1. Benchmarking against the practice of corporate tax avoidance**

In exploratory analysis, we look at the reallocation decisions of spectators in a tax avoidance scenario (figure 4). Since this kind of activity is better known, and broadly viewed as unethical, it was useful for us to benchmark our results from the primary analysis against. This could help us say how much weight spectators assign, on average, to the use of algorithms to assess applications relative to a company avoiding taxes. In the control condition, the difference between how much people reallocate in the tax avoidance scenario compared to the prompt scenario is 127.03 pence (with a standard deviation of 267.46p). In

other words, compared to a neutral description of data use, spectators punish tax avoidance behaviour by 127p more (more than a fourfold increase). We know from the primary analysis that the increase in money moved in the treatment conditions is 34.6 - 45.2 pence, depending on the specific prompt. This means that spectators “punish” aggressive use of algorithms by deducting 27.2-35.6% as much revenue as they would do if presented with a tax avoidance behaviour instead of algorithm use. In other words, on average people weigh the aggressive use of algorithms about one third as much as they would a company avoiding taxes.



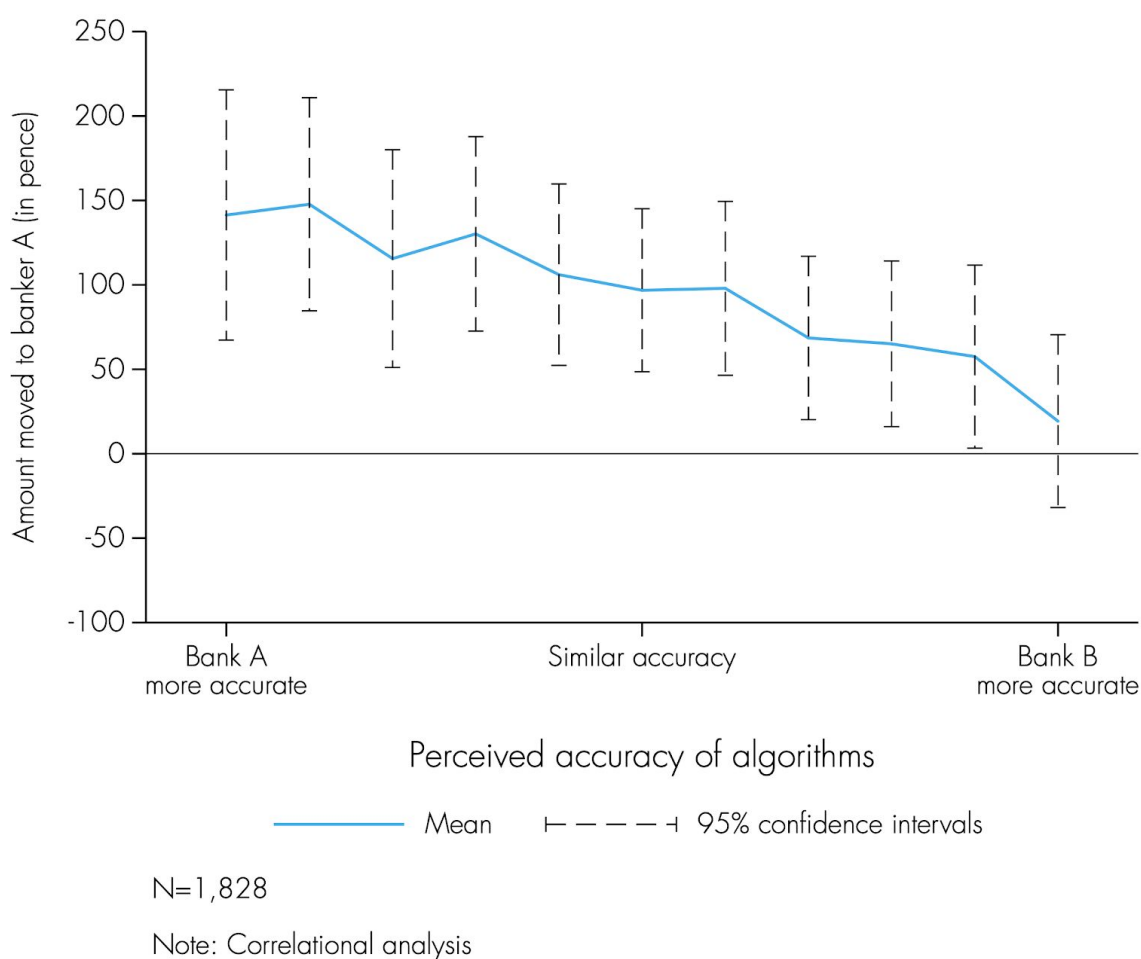
**Figure 4: The comparison between reallocation in the tax avoidance algorithm use scenarios**

Note that we calculated the proportion by looking at the reallocation in the tax avoidance scenario in the control condition. Interestingly, the amount reallocated in the tax avoidance scenario is lower in the treatments compared to the control (-26p in the gender prompt (p-value=0.123); -15p in the ethnicity prompt (p-value=0.362) and -45.9p in the social media prompt (p-value=0.009)). This suggests that people could be influenced in their decision by the previous scenario. Specifically, judging a respondent more harshly in the first scenario (as spectators in the treatment do) can result in a less harsh judgment in the following scenario. This is in line with an oscillating pattern of moral judgment.<sup>18</sup> A second caveat with this analysis is that the tax avoidance scenario was not incentivised due to budgetary constraints. This could cause punishment levels to be higher in the tax avoidance scenario than they would be in an incentivised elicitation.

<sup>18</sup> Gneezy, U., Imas, A., & Madarász, K. (2014). Conscience accounting: Emotion dynamics and social behavior. *Management Science*, 60(11), 2645-2658.

A second interesting finding in this analysis is that demographic characteristics are strongly correlated with the amount that is reallocated in the tax avoidance scenario (recall that we found no such correlations in the algorithm use scenario). Specifically, older individuals and those with higher educational attainment reallocate more. These regression results can be found in the appendix.

### 2.2.5.2. The relevance of perceived accuracy



**Figure 5: The correlation between perceived accuracy and allocation to Banker A**

Table 11 reports the regression results of the reallocated amount by spectators when controlling for perceived accuracy of advanced computing techniques by Bank B. Figure 5 displays these results graphically. There is a significant negative correlation between higher perceived accuracy of these techniques and the amount that is moved towards banker A. This suggests that individuals who believe that bank B's computing techniques are more accurate are less likely to perceive the use of these techniques as unfair. A regression with interaction terms by treatment suggests that this holds even when individuals are prompted that these techniques might have adverse consequences.



**Table 11: Regression results of reallocated amount including control for perceived accuracy**

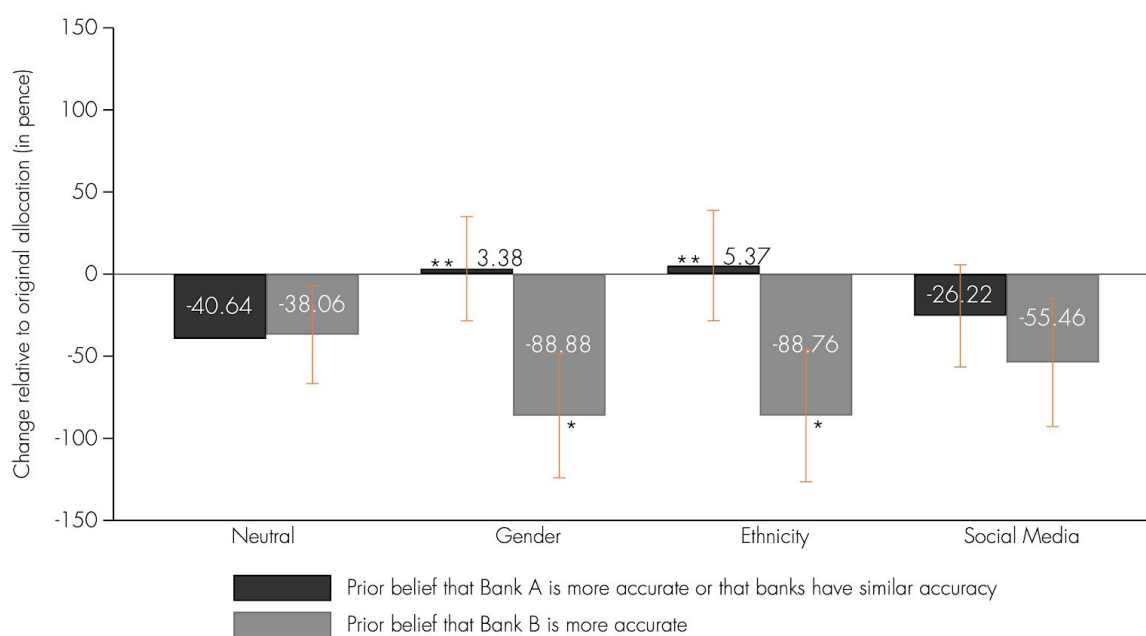
	Amount reallocated (pence)
<b>Treatment</b> (baseline = Neutral)	
Gender prompt	30.209** (11.27)
Ethnicity prompt	29.268* (12.10)
Social Media prompt	37.628** (11.68)
Perceived accuracy of advanced computing techniques (11-point scale)	-12.588** (2.78)
Demographics controls	YES
Constant	62.980* (31.60)
Observations	1,828
R-squared	0.06
Robust standard errors in parentheses; ** p<0.01, * p<0.05, + p<0.1	

Finally, we look at the change in allocation when spectators are instructed to assume that Bank B's computing techniques allow them to make more accurate predictions about the applicant's reliability than Bank A. They were then asked if they wanted to change their allocation relative to what they had decided originally (this decision is hypothetical and did not affect the payment to the bankers). Table 12 presents the regression results where the outcome variable is the difference between the new allocation and the old allocation. Negative numbers indicate that the spectator allocates less to banker A when they are assuming that Bank B's techniques are more accurate compared to their original allocation. In this analysis, we have split the sample according to the spectator's perceived accuracy beliefs. We create one group that believes that Bank A is more accurate or that the banks have similar accuracy. The other group believes that Bank B is more accurate. Note that our results hold when we include perceived accuracy as a continuous variable in the regression. Figure 6 displays these results graphically.

**Table 12: Regression results - change in allocation when assuming higher accuracy for Bank B**

	Change in allocation when assuming higher accuracy (in pence)
<b>Treatment</b> (baseline = Neutral)	
Gender prompt	42.734** (16.17)
Ethnicity prompt	44.664** (17.12)

Social Media prompt	14.002 (15.86)
Prior belief that Bank B is more accurate than Bank A	2.504 (15.14)
Interaction (Gender prompt * Prior belief that Bank B is more accurate than Bank A)	-46.836* (19.29)
Interaction (Ethnicity prompt * Prior belief that Bank B is more accurate than Bank A)	-46.711* (20.55)
Interaction (Social Media prompt * Prior belief that Bank B is more accurate than Bank A)	-14.387 (19.89)
Demographic controls	YES
Constant	-39.458 (24.74)
Observations	1,828
R-squared	0.025
Robust standard errors in parentheses; ** p<0.01, * p<0.05, + p<0.1	



N=1,828

Note: Exploratory analysis

**Figure 6: The change in allocation relative to the original decision when the spectator believes Bank B to be more accurate, conditional on prior beliefs**

Spectators who have a prior belief that Bank A is more accurate or that the banks have similar accuracy reduce their allocation to Bank A by 39p when asked to assume that Bank B is more accurate. Interestingly, this adjustment only happens in the control conditions and the social media conditions. In the gender and ethnicity prompts this reduction does not happen - spectators keep roughly the same allocation as in their original decision, even though Bank B is now more accurate. For individuals who already thought Bank B was more accurate we see a large reduction across the board, though this effect is not statistically significant in the social media condition. This analysis creates some more nuance around a possible causal relationship between communicating accuracy and fairness perceptions. In particular, for those who currently are not as convinced about the accuracy of algorithms, the results suggest that such information is most influential when their use is described in neutral terms or relates to social media information.

## 2.3 Conclusion

As expected, people react to the use of information which could act as a proxy for other characteristics by banks making loan decisions. We see relatively consistent aversion to them compared to a neutral control such that they financially punish banks that are known to use such techniques. In the real world, this could result in lower take-up of financial products, consumers switching to other providers or sharing unfavourable reviews about the bank, for example. However, switching rates in general are known to be low,<sup>19</sup> so consumers might not get round to acting even if they are unhappy. In addition, financial providers may be (unintentionally) incentivised not to be transparent about their practices to avoid any risk of losing business. As we discuss below, the proportion of consumers reacting to information related to algorithmic decision-making is likely to be sensitive to the framing of the way in which information may act as a proxy.

Although we cannot be certain at this point due to the limited sample sizes (particularly for ethnic minority participants), the directional effects of the results suggests that people perceived to be disadvantaged by the algorithmic decision-making, women and ethnic minorities, are more likely to punish the banks than others. This is an important consideration for policymakers as these groups may represent a group with less power within financial services structures.

People's views of social media use in algorithmic decision-making is particularly interesting. Although the differences between conditions is not statistically significant, directionally the results suggest that people feel most uncomfortable with the potential use of social media as a decision metric.

While people do indeed react to the use of information which could act as a proxy for other characteristics, this is equivalent to about a third of their aversion to tax avoidance - considerable but not overwhelming dissatisfaction. This research also suggests that even

---

<sup>19</sup> See Citizens Advice research on the [Loyalty Penalty](#).

minor changes to wording can shift perceptions. For example, in our exploratory research, when people were exposed to the treatment conditions they punished banks who avoided tax less than those in the control condition despite the scenarios being identical. This was only statistically significant in the social media condition. This suggests that when people are primed with potentially unethical behaviour by banks they are more forgiving of other unethical behaviours.

Finally, we see that the framing and ordering of information influences perceptions of fairness. Firstly, when people are told that the algorithm is more accurate, they perceive its use more favourably. In theory, financial institutions could say that their algorithm is more accurate, with little evidence of accuracy, or overclaim its accuracy (particularly where such practices are not currently audited). Conversely, this information may be valuable for consumers in supporting the use of sophisticated technologies which act to increase equality. Secondly, we see that people exposed to the treatment conditions (i.e., the proxy banks) punish the bank that avoids paying tax less than those in the neutral control condition, despite the fact that there is no difference between the questions that they were asked. It may be that when contentious practices by banks are considered cumulatively, people assume they are the norm and adjust their levels of acceptability.

## 2.4 Recommendations

Below we primarily discuss recommendations related to improving the transparency of algorithmic decision-making reflecting our research results. However, we acknowledge the limitations of transparency as a means of solving all policy challenges. What we do not cover in-depth below is where financial institutions should instead not use such methods, particularly where groups are discriminated against as a result of proxy information. We welcome further discussion on this.

### 2.4.1 Recommendations for the CDEI

Perceptions of algorithm accuracy affects how fair people perceive its use. More needs to be done to translate what this means in practice for consumers. If people are willing to tolerate differences in outcome when the algorithm is understood to be more accurate, this could suggest that financial institutions should be more open about how accurate their algorithms are. However, we should also consider how consumers understand the measure of accuracy and what it means in reality. The CDEI should consider further testing to understand the impact of framing on perceptions of fairness, acceptance and comprehension of algorithm use, for example:

1. Vary the framing of the use of information which may act as a proxy;
2. Vary how algorithm accuracy is framed;
3. Consider their interaction with each other.

In addition, this should look to unpick whether the claim of greater accuracy would be treated with more scepticism by those who believe they are most at risk of being disadvantaged.

On use of social media data, we recommend that the CDEI:

1. Collate a clearer understanding of its current or potential use by financial services, among other sectors;
2. Provide consumers with this information and further explore how they perceive its use. For example, with more information might we see clearer differences between those who use social more frequently, something that we do not see in this research;
3. Ensure policy either protects or empowers consumers to get the most appropriate outcome (relevant to broader policy recommendations).

#### **2.4.2 Recommendations for policy makers and financial services**

4. Encourage financial institutions who use algorithms to screen customers to test whether their systems may be (inadvertently) biased on the basis of gender or ethnicity, as a result of other information that acts as a proxy for the characteristics. This is important as our research shows that people perceive this as an unfair practice.
5. Consider whether policies to improve transparency will, on their own, be sufficient, given that those most likely to be affected by this issue perceive it as most unfair. It may be that financial services respond to customer dissatisfaction and reduce their discriminatory practices given the proportion of customers affected and their response to the information. However, this may not be a sufficient incentive for financial services to change their practices and more may need to be done to protect those affected consumers (e.g., through stronger regulation, increased monitoring).
6. All financial institutions should be using sophisticated algorithms to better serve their customers, while considering and monitoring the outcomes for different groups such as women and ethnic minorities. This should also incorporate the use of deliberative forums for consumers to feedback directly on its use.

# Appendices

---

## Appendix 1: Full participant instructions and treatment materials

*All materials can be found on the Predictiv dashboard:*

- [Experiment for the bankers](#)
- [Experiment for the spectators](#)

## Appendix 2: Power Calculation Code

```

sample.size <- c(400, 450, 500, 600)

# effect sizes in Cohen's D
effect.sizes <- c(0.1, 0.125, 0.15, 0.175, 0.2, 0.3, 0.4)

sds <- seq(from=10, to=150, by=20)

lengthy <- length(effect.sizes) * length(sample.size) * length(sds)

cohensD <- rep(NA, lengthy)
ns <- rep(NA, lengthy)
ns.per.arm <- rep(NA, lengthy)
pw <- rep(NA, lengthy)
a <- rep(NA, lengthy)
effect <- rep(NA, lengthy)
truediffm <- rep(NA, lengthy)
cohensD <- rep(NA, lengthy)

r <- 0

pc <- data.frame(cbind(ns, ns.per.arm, truediffm, cohensD, pw, a))

library(pwr)

for (j in 1:length(sample.size)){
  for (i in 1:length(effect.sizes)){
    for (k in 1:length(sds)){

      r <- r+1

      hold <- power.t.test(n = sample.size[j], d =effect.sizes[i], sig.level = 0.05,
                           power = NULL,
                           type = "two.sample",
                           alternative = "two.sided")

      pc$ns[r] <- sample.size[j] * 4 # as there are 4 trial arms
      pc$ns.per.arm[r] <- sample.size[j]
      pc$cohensD[r] <- effect.sizes[i]
      pc$truediffm[r] <- sds[k] * effect.sizes[i]
      pc$sd[r] <- sds[k]
      pc$pw[r] <- hold$power
      pc$a[r] <- 0.05

    }
  }
}

write.csv(pc, "TP2019018_PowerTable1.csv")

pc <- NULL

### Test 3.1a: amount reallocated (as raw number, between -500 and +650) -- power held to 80% and solving for
MDES

```

```

# solving for effect sizes in Cohen's D

sample.size <- seq(from=300, to=800, by = 50)

sds <- seq(from=10, to=150, by=20)

lengthy <- length(sample.size) * length(sds)

cohensD <- rep(NA, lengthy)
ns <- rep(NA,lengthy)
ns.per.arm <- rep(NA,lengthy)
pw <- rep(NA,lengthy)
a <- rep(NA,lengthy)
effect <- rep(NA,lengthy)
truediffm <- rep(NA,lengthy)
cohensD <- rep(NA,lengthy)
b <- rep(NA,lengthy)

r <- 0

pc <- data.frame(cbind(ns, ns.per.arm, truediffm, cohensD, pw, a))

library(pwr)

for (j in 1:length(sample.size)){
  for (k in 1:length(sds)){

    r <- r+1

    hold <- power.t.test(n = sample.size[j], d=, sig.level = 0.05,
                        power = 0.8,
                        type = "two.sample",
                        alternative = "two.sided")

    pc$ns[r] <- sample.size[j] * 4 # as there are 4 trial arms
    pc$ns.per.arm[r] <- sample.size[j]
    pc$cohensD[r] <- hold$d
    pc$truediffm[r] <- sds[k] * hold$d
    pc$sd[r] <- sds[k]
    pc$pw[r] <- hold$power
    pc$a[r] <- 0.05
    pc$b[r] <- 0.8

  }
}

write.csv(pc, "TP2019018_PowerTable2.csv")

```



## Appendix 3: Additional results

**Table A1: Demographic breakdown of the sample by treatment**

	Neutral (%)	Gender (%)	Ethnicity (%)	Social media (%)
<b>Gender</b>				
Female	49.6	47.3	46.5	51.2
Male	50.4	52.7	53.5	48.8
<b>Age</b>				
18-24 years	25.2	29.0	26.9	29.2
25-39 years	28.7	22.9	22.9	25.9
40 - 54 years	20.2	20.7	22.7	21.4
55 and over	25.9	27.5	27.5	23.5
<b>Income</b>				
Below median (<=£27,499)	49.1	49.9	50.0	51.4
Above median (>=£27,500)	50.9	50.1	50.0	48.9
<b>Location</b>				
London	13.8	14.8	15.9	14.4
North	24.8	24.4	21.8	22.7
South & East	28.7	31.2	34.6	29.4
Midlands	19.1	16.1	13.9	14.4
Wales, Scotland & N Ireland	13.6	13.5	13.9	16.8
<b>Education (highest)</b>				
None	0.9	1.1	1.1	2.4
GSCE	20.0	21.6	22.5	19.4
A-level	20.2	20.0	21.2	19.6
Vocational diploma	16.2	15.0	15.4	17.2
Undergraduate	27.0	27.2	23.6	29.0
Professional qualification	6.4	5.2	4.9	2.6
Postgraduate	9.4	9.8	11.5	9.8
<b>Ethnicity</b>				
White	86.8	87.8	87.7	86.1
Mixed/multiple ethnic groups	2.4	1.7	2.9	2.8
Asian/Asian British	6.1	7.2	4.6	6.3
Black/African/Caribbean/Black British	4.4	2.6	3.7	4.1
Other	0.2	0.7	1.1	0.7
<b>Social media use - platform #</b>				
None	8.3	9.6	9.7	9.2
1	22.2	21.6	21.6	20.5

2-3	49.1	42.3	42.5	42.5
4-6	18.2	24.8	24.0	24.0
7 or more	2.2	1.7	2.2	3.9
<b>Social media use - frequency</b>				
Never	7.7	8.7	8.8	9.4
Once a week	8.3	8.7	8.6	10.7
Couple of times a week	12.3	7.8	8.2	9.4
Once a day	11.8	10.9	15.2	9.8
Couple of times a day	24.6	21.4	21.6	19.6
Many times a day	35.3	42.5	37.7	41.2
<b>Computer skills</b>				
1 - difficulty understanding	0.7	1.1	0.4	1.3
2	2.6	2.2	3.1	1.7
3 - comfortable	23.0	25.9	29.5	23.8
4	26.1	25.1	24.2	26.4
5 - very comfortable	47.6	45.8	42.7	46.8
<b>Previous loan experience</b>				
Has applied for a loan	48.9	47.5	45.2	47.7
Has not applied for a loan	46.1	50.1	52.0	49.2
Don't know/prefer not to say	5.0	2.4	2.8	3.1

**Table A2: Secondary regression results - gender subgroup effects**

	Amount reallocated (pence)
<b>Treatment</b> (baseline = non-Gender prompts)	
Gender prompt	-4.7 (14.40)
Female (baseline is male)	1.921 (9.14)
Interaction (Gender prompt * Female)	25.847 (18.60)
Demographic controls	YES
Constant	67.745* (30.92)
Observations	1,828
R-squared	0.015
Robust standard errors in parentheses; ** p<0.01, * p<0.05, + p<0.1. Note that our treatment variable is defined here as a binary variable with a value of 1 if the participant is in the gender treatment, and 0 otherwise.	

**Table A3: Secondary regression results - ethnicity subgroup effects**

	Amount reallocated (pence)
<b>Treatment</b> (baseline = non-Ethnicity prompts)	
Ethnicity prompt	0.317 (29.03)
White ethnicity (baseline is non-white)	-9.994 (15.22)
Interaction (Ethnicity prompt * White ethnicity)	9.041 (30.82)
Demographic controls	YES
Constant	79.743* (33.03)
Observations	1,828
R-squared	0.01
Robust standard errors in parentheses; ** p<0.01, * p<0.05, + p<0.1. Note that our treatment variable is defined here as a binary variable with a value of 1 if the participant is in the ethnicity treatment, and 0 otherwise.	

**Table A4: Secondary regression results - social media subgroup effects**

	Model 1: Platform frequency	Model 2: Age bracket
	Amount reallocated (pence)	Amount reallocated (pence)
<b>Treatment</b> (baseline = non-Social Media prompts)		
Social Media prompt	39.336 (42.90)	-5.046 (15.59)
<b>Frequency of social media use</b>		
Once a week	-5.351 (21.72)	
Couple time a week	-19.461 (22.09)	
Once a day	-6.422 (22.68)	
Couple times a day	14.882 (21.27)	
Many times a day	-1.407 (20.56)	
<b>Interaction (Social Media prompt * Frequency of social media use)</b>		
Once a week	-33.648 (51.77)	
Couple times a week	15.13 (54.05)	
Once a day	-7.976	

	(48.88)	
Couple times a day	-28.364 (47.19)	
Many times a day	-20.84 (45.29)	
<b>Age bracket</b> (baseline = 18-24 years)		
25-39 years		-16.068 (12.74)
40-54 years		-16.702 (13.68)
55+ years		-4.723 (12.73)
<b>Interaction (Social Media prompt * Age bracket)</b>		
25-39 years		46.215+ (24.92)
40-54 years		66.999* (29.47)
55+ years		2.34 (23.16)
Demographic controls	YES	YES
Constant	59.284+ (34.42)	66.987* (29.03)
Observations	1,828	1,828
R-squared	0.02	0.021

Robust standard errors in parentheses; \*\* p<0.01, \* p<0.05, + p<0.1. Note that our treatment variable is defined here as a binary variable with a value of 1 if the participant is in the social media treatment, and 0 otherwise.

**Table A5: Regression results of reallocated amount in the tax avoidance scenario.**

	Amount reallocated (pence)
<b>Treatment</b> (baseline = Neutral)	
Gender prompt	-26.414 (17.12)
Ethnicity prompt	-15.274 (16.76)
Social Media prompt	-45.919** (17.43)
Gender (baseline = Male)	-8.92 (12.36)
<b>Age bracket</b> (baseline = 18-24 years)	

25-39 years	-5.498 (16.60)
40-54 years	35.553+ (18.30)
55+ years	70.753** (16.95)
<b>Household income bracket</b> (baseline = £27,499 or lower)	
£27,500 and above	25.412* (12.91)
<b>Location</b> (NUTS; baseline = London)	
North	18.673 (20.34)
South & East	34.740+ (19.70)
Midlands	8.327 (21.99)
Wales, Scotland & N Ireland	48.862* (23.35)
<b>Ethnicity</b> (baseline = White)	
Mixed/multiple ethnic groups	-0.291 (34.73)
Asian / Asian British	2.055 (22.21)
Black/African/Caribbean/Black British	-16.294 (27.55)
Other	-42.189 (64.04)
<b>Education</b> (highest obtained; baseline = None)	
GSCE	61.142+ (36.07)
A-level	93.274** (35.69)
Vocational diploma	74.335* (36.74)
Undergraduate	114.401** (35.62)
Prof qual	80.500+ (43.78)
Postgraduate	57.202 (39.56)
Constant	16.246 (39.24)

Observations	1,828
R-squared	0.031
Robust standard errors in parentheses; ** p<0.01, * p<0.05, + p<0.1	