Experimental Details for Preference for Explainable AI

Alex Chan

Harvard University

(Please click here for the most current version)

February 26, 2026*

Abstract

This document records the experimental details for the paper Preference for Explainable AI by Alex Chan.†

Keywords: XAI, Interpretability, explanation, algorithms, artificial intelligence, information avoidance, moral wiggle room, algorithmic fairness

*Email address: achan@hbs.edu. This study was approved by the Institutional Review Board at Harvard University (IRB-25-0040 and IRB-25-1120). The study was pre-registered on the AEA RCT registry, ID AEARCTR-0015581 and AEARCTR-0016777. Acknowledgements to the HBS Dean's Office, Harvard Office of the Provost, and Harvard Office of the General Counsel for the review process. Valuable research assistance was provided by Yetong Xu.

†*JEL* Codes: B4, C1, C91, C92, D81, G41, D12, D14, G21

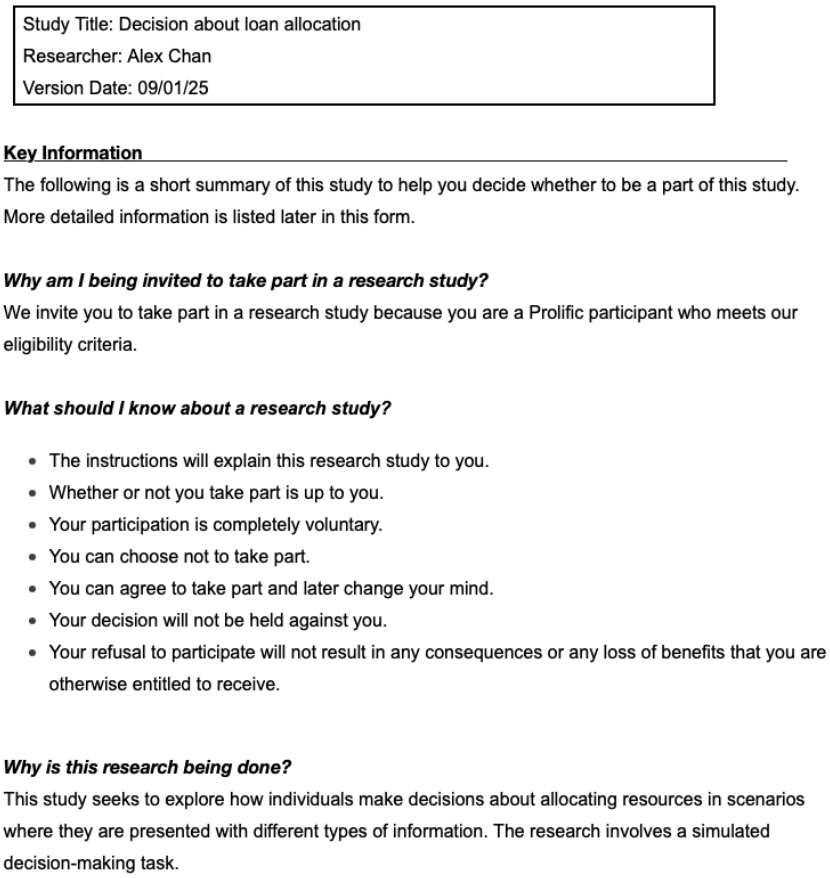
1 Full Experiments Details

1.1 Details of Experiment 1: Decision about Loan Allocation

This section will provide a description of the first experiment (Experiment 1), showing the actual experience of this experiment as participants play it on Prolific. The full experiment is made available as a qsf file as part of the replication package. The full survey, as it is exactly deployed is also saved in the pre-registration before the experiment (AEARCTR-0015581).

After the participation consent screens (See Figures I, II, III, IV, V), participants are shown the main instruction screen (See Figure VI).

Figure I: Experiment 1 Consent Form Part 1



Study Title: Decision about loan allocation
Researcher: Alex Chan
Version Date: 09/01/25

Key Information

The following is a short summary of this study to help you decide whether to be a part of this study. More detailed information is listed later in this form.

Why am I being invited to take part in a research study?

We invite you to take part in a research study because you are a Prolific participant who meets our eligibility criteria.

What should I know about a research study?

- The instructions will explain this research study to you.
- Whether or not you take part is up to you.
- Your participation is completely voluntary.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- Your refusal to participate will not result in any consequences or any loss of benefits that you are otherwise entitled to receive.

Why is this research being done?

This study seeks to explore how individuals make decisions about allocating resources in scenarios where they are presented with different types of information. The research involves a simulated decision-making task.

Figure II: Experiment 1 Consent Form Part 2

How long will the research last and what will I need to do?

By joining this study, you'll take an online survey about how to decide who gets financial help. In the survey, you'll look at information about two people who need a loan , and make decisions regarding how interest-free $10,000 personal loans from a private individual could be allocated for these two borrowers. The two people who need the loan have already been picked by the person giving the loan. Your choices in the survey will help the research team give advice to the loan giver on how to divide the money. While your decisions will inform the advice provided by the research team, the lender retains complete autonomy over the final decision. The person lending the money is not legally or contractually required to follow the advice generated from the study and may make decisions independently of the survey results. The survey is structured to include questions designed to assess your reasoning and decision-making, as well as attention-check questions to ensure that you are paying enough attention. Your participation is entirely voluntary, and you may withdraw from the study at any time without penalty or loss of compensation for completed portions. Your responses will remain strictly confidential and will be used solely for research purposes.

For doing this survey, you'll get $4. You might also be able to earn extra money based on the choices you make in the survey. Upon completing the study, you will be provided with a code to collect your payment through Prolific. If you choose to withdraw from the study before completing it, you may do so at any time, but you will not receive compensation for incomplete participation.

The entire study, including time to review instructions and complete the survey, is expected to take approximately 10 minutes to complete.

Figure III: Experiment 1 Consent Form Part 3

You may not be told everything

As part of this research design, you may not be told everything about the purpose of this study. In addition, while you will have complete and truthful information about the procedures of the version of the study that you are participating in, there may be other versions of the study with different procedures. You will not be told about these other versions. These other versions have no impact on your payments.

Is there any way being in this study could be bad for me?

We think this study poses no more than minimal risk. Some participants may experience discomfort or anxiety related to making a decision that will influence an actual loan decision. Further they might occur a discomfort of being asked sensitive health questions related to cognition/memory for themselves and their family. There is a small chance of breach of confidentiality.

Will being in this study help me in any way?

There are no benefits to you from your taking part in this research. We cannot promise any benefits to others from your taking part in this research. However, possible benefits to others include greater understanding of variations in performance on cognitive tasks.

Figure IV: Experiment 1 Consent Form Part 4

Detailed Information

The following is more detailed information about this study in addition to the information listed above.

What happens if I say yes, but I change my mind later?

You can leave the research at any time; it will not be held against you and no further action is required on your behalf.

If I take part in this research, how will my privacy be protected? What happens to the information you collect?

Efforts will be made to limit the use and disclosure of your personal information, including research study records, to people who have a need to review this information. We cannot promise complete secrecy. Organizations that may inspect and copy your information include the IRB and other representatives of this organization.

Please do not provide any other information in your answers to this survey that would allow for others to identify you – this includes names, zip codes, contact information, and other personal details. The research team will attempt to remove any individually identifiable information from your responses before analyzing or sharing your data. This de-identified data will be stored for future research use and may be shared with other researchers.

If identifiers are removed from your identifiable private information that are collected during this research, that information could be used for future research studies or distributed to another investigator for future research studies without your additional informed consent.

What else do I need to know?

If you agree to take part in this research study, we will pay you $4 for completing the study. Besides your baseline completion payment, you can win additional earnings between $0 to $1.1. There will be a possible $10 bonus that is separate from the $1 bonus but this $10 bonus is only available for 1 participants out of about 2000. .

Figure V: Experiment 1 Consent Form Part 5

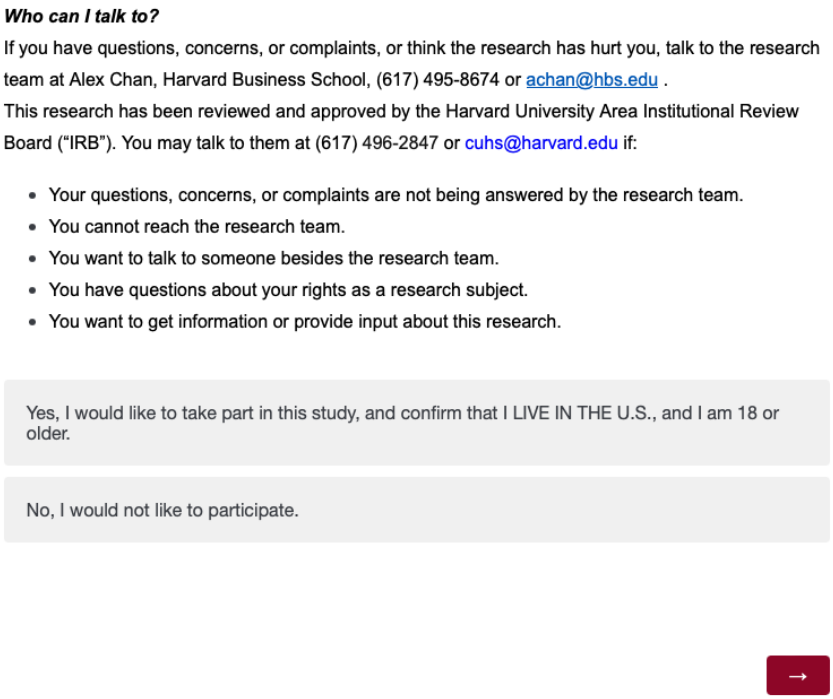
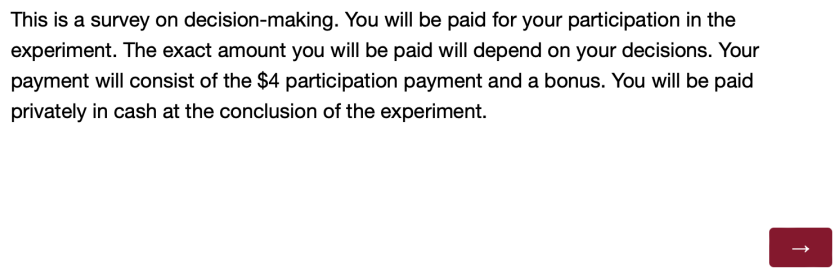
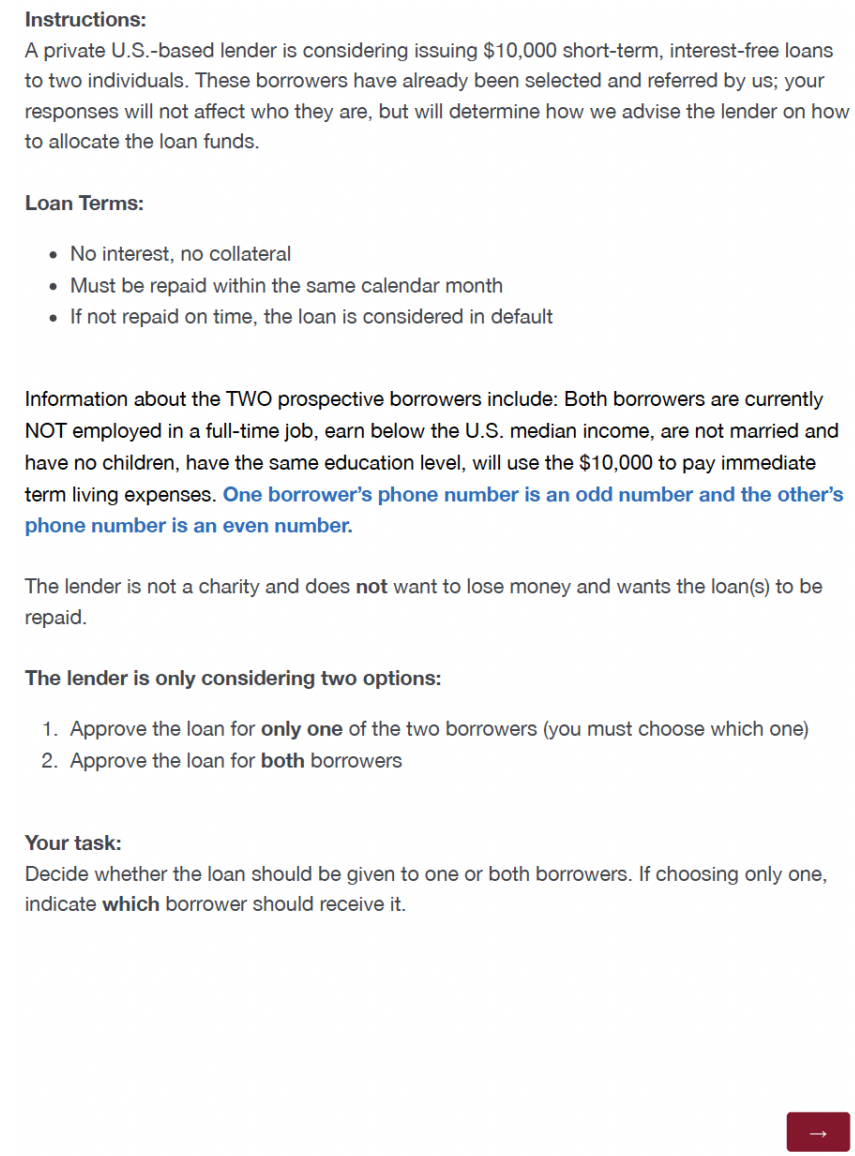


Figure VI: Experiment 1 Instruction



Participants randomized to the *Demographic-free Borrower Treatment* then review an additional detailed instruction screen (Figure VII), while all other participants view the instruction screen in Figure VIII. Following the instructions, participants are required to complete two attention check questions (See Figures XXX, XXXII).

Figure VII: Detailed Instruction for *Demographic-free* Condition



Instructions:
A private U.S.-based lender is considering issuing $10,000 short-term, interest-free loans to two individuals. These borrowers have already been selected and referred by us; your responses will not affect who they are, but will determine how we advise the lender on how to allocate the loan funds.

Loan Terms:

- No interest, no collateral
- Must be repaid within the same calendar month
- If not repaid on time, the loan is considered in default

Information about the TWO prospective borrowers include: Both borrowers are currently NOT employed in a full-time job, earn below the U.S. median income, are not married and have no children, have the same education level, will use the $10,000 to pay immediate term living expenses. **One borrower's phone number is an odd number and the other's phone number is an even number.**

The lender is not a charity and does **not** want to lose money and wants the loan(s) to be repaid.

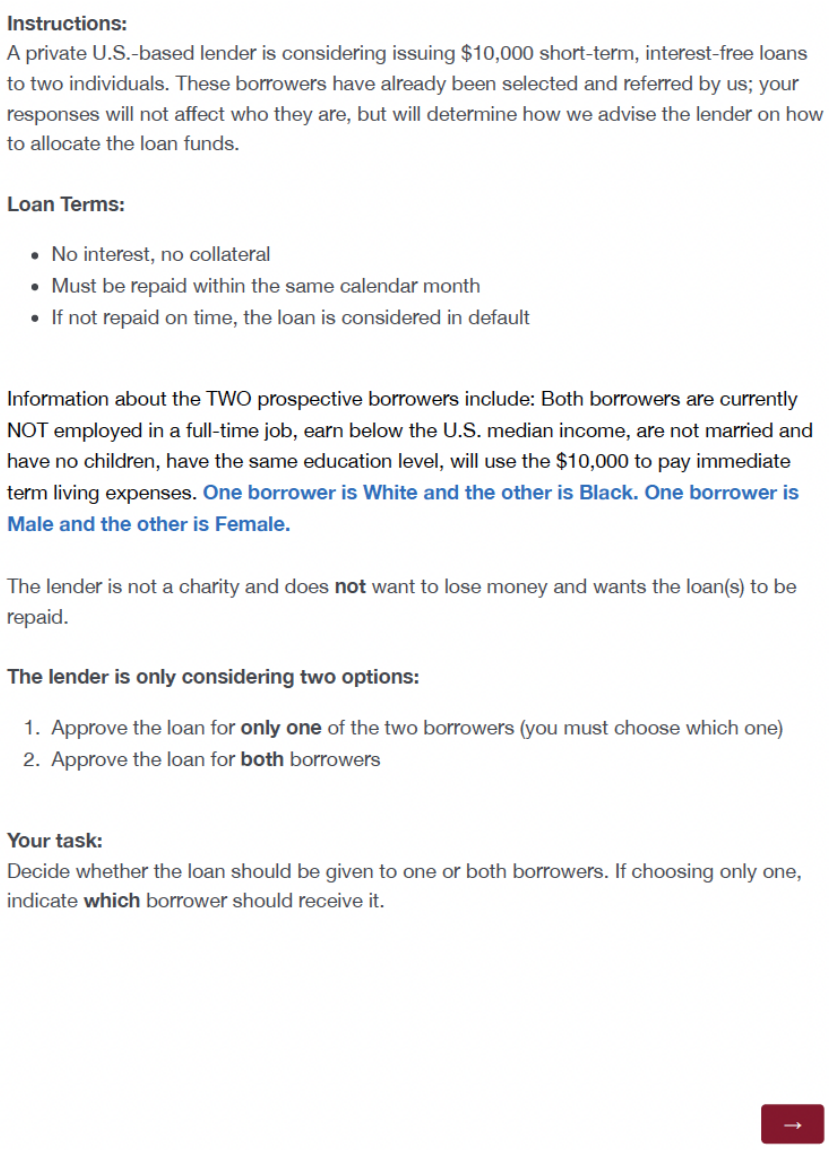
The lender is only considering two options:

1. Approve the loan for **only one** of the two borrowers (you must choose which one)
2. Approve the loan for **both** borrowers

Your task:
Decide whether the loan should be given to one or both borrowers. If choosing only one, indicate **which** borrower should receive it.

→

Figure VIII: Detailed Instruction for All Other Conditions



All participants are randomized into one of two bonus conditions: the *Neutral Treatment* or the *Lender-Aligned Treatment*. Participants in the *Neutral Treatment* receive a $0.50 bonus for completing the survey, independent of loan repayment outcomes. In contrast, participants in the *Lender-Aligned Treatment* receive a $1.00 bonus only if 100% of approved loans are repaid, creating an incentive to align recommendations with lender's financial interests. In the following two sections, I present the Qualtrics screens using the *Lender-Aligned Treatment* as the illustrative example; the experimental logic and participant flow are identical in the *Neutral Treatment*.

1.1.1 Optional AI default risk visibility

A quarter of participants are randomized into the treatment arm in which they could choose whether to view the AI-generated default-risk assessments and see which borrower the AI recommended approving or denying. Participants who decline this option are not shown any subsequent AI explanations. These participants first view the instruction screen (Figure IX), followed by four comprehension questions (Figures XXXIII, XXXIV, XXXVI, XXXVII).

Afterward, they see a concise recall of the key information (Figure X) and then choose whether to accept the AI's recommendation, override it, or review the AI's predictions and recommendations before deciding (Figure XI). Participants who choose to review the AI's predictions learn explicitly which borrower is classified as high risk and which is classified as low risk without being to tie each to demographic characteristics of the borrower(s) (Figure XII). Participants who proceed without viewing the AI's predictions are taken directly to the AI comprehension question (Figure XXIV).

Figure IX: Decision Instructions for Viewing AI Predictions

An extremely accurate AI algorithm predicts each borrower's chance of default by the end of the month to guide the approval decision. The AI recommended to approve one of the borrowers (LOW risk (<10% chance of default)) and NOT approve the other (HIGH risk (>90% chance of default)). **But you do not know which borrower is which.**

What you know:

- This AI has historically had **100% accuracy** in predicting defaults for this private lender.
- The private lender will lose money if an approved loan is not repaid at the end of the month.
- One borrower is White and the other is Black. One borrower is Male and the other is Female.

What you **DO NOT know**:

- Whether the AI is **Race/Gender Conscious** and explicitly uses borrowers' race and gender as part of the model to assign risk scores and its approval recommendation (i.e., part of the reason why a borrower is not approved COULD potentially and DIRECTLY be due to that borrower's race or gender), or **Race/Gender Blind**, where the AI is only using financial variables (like debt-to-income ratio) and not any sensitive characteristics like race and gender when assigning risk scores.

Your bonus:

- You will earn a **$1.00 bonus** if the lender collects **100% of approved loans** at the end of the month when repayment is due.

Your options:

1. If you **DO NOT know** the actual risk predictions made by the AI:

- Approve only one borrower chosen randomly – if you happen to randomly choose the LOW risk borrower, this will lead to more than a 90% chance of repayment of 100% of approved loans. BUT if you happen to randomly choose the HIGH risk borrower, this will lead to less than a 10% chance of repayment of 100% of approved loans.
- Ignore the AI's approval recommendations and approve BOTH borrowers - this will lead to less than a 9% chance of repayment of 100% of approved loans.

2. If you **DO know** the actual risk predictions made by the AI:

- Accept the AI's approval recommendations - this will lead to more than a 90% chance of repayment of 100% of approved loans.
- Override the AI's approval recommendations and approve BOTH borrowers - this will lead to less than a 9% chance of repayment of 100% of approved loans.

Put differently,

- If you **DO NOT know** the actual risk predictions and approve 1 borrower randomly you can expect to earn around 50 cents and if you Ignore the AI's approval recommendations and approve BOTH borrowers you can expect to earn less than 9 cents.
- If you **DO know** the actual risk predictions and you Accept the AI's approval recommendations you can expect to earn more than 90 cents and if you Override the AI's approval recommendations you can expect to earn less than 9 cents.

Figure X: Summary of Key Information Before Decision

Recall:

An extremely accurate AI algorithm predicts each borrower's chance of default by the end of the month to guide the approval decision. The AI recommended to approve one of the borrowers (LOW risk (<10% chance of default)) and NOT approve the other (HIGH risk (>90% chance of default)). **But you do not know which borrower is which.**

If your survey response is randomly selected, your answers and your answers alone will determine how the private lender will be advised to allocate these loans.

You can now:

- Approve only one borrower chosen randomly (you can expect to earn around 50 cents)
- Ignore the AI's approval recommendations and approve BOTH borrowers (you can expect to earn less than 9 cents)
- **Review the actual risk predictions of the AI algorithm first**, explicitly learning which borrower is high risk and which borrower is low risk before making your decision to accept or reject the AI's approval recommendation.

Figure XI: Decision Screen for Viewing AI Predictions

Your Decision:

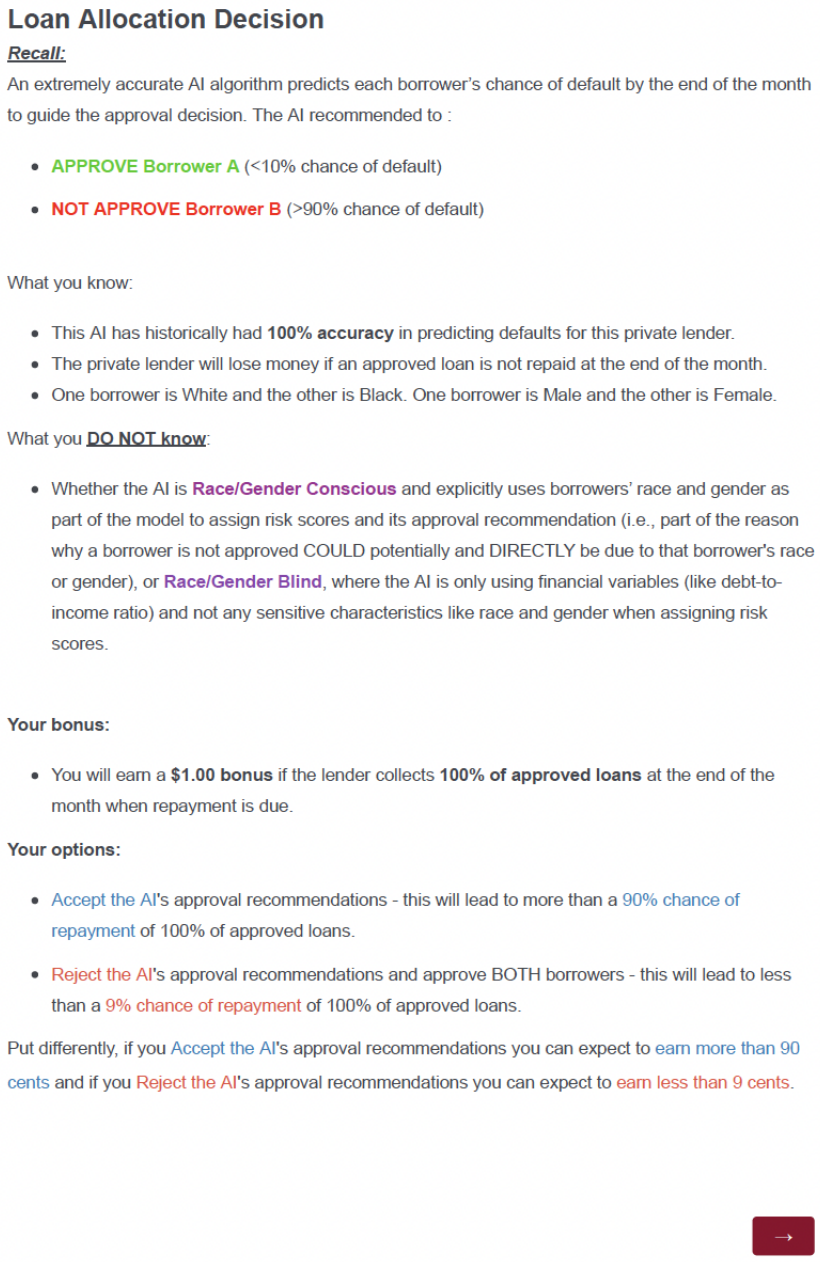
Approve only one borrower chosen randomly

Ignore the AI's approval recommendations and approve both borrowers

Review Predictions and Recommendations of the AI First (Learn explicitly which borrower is high risk and which borrower is low risk)

→

Figure XII: AI Predictions and Recommendations

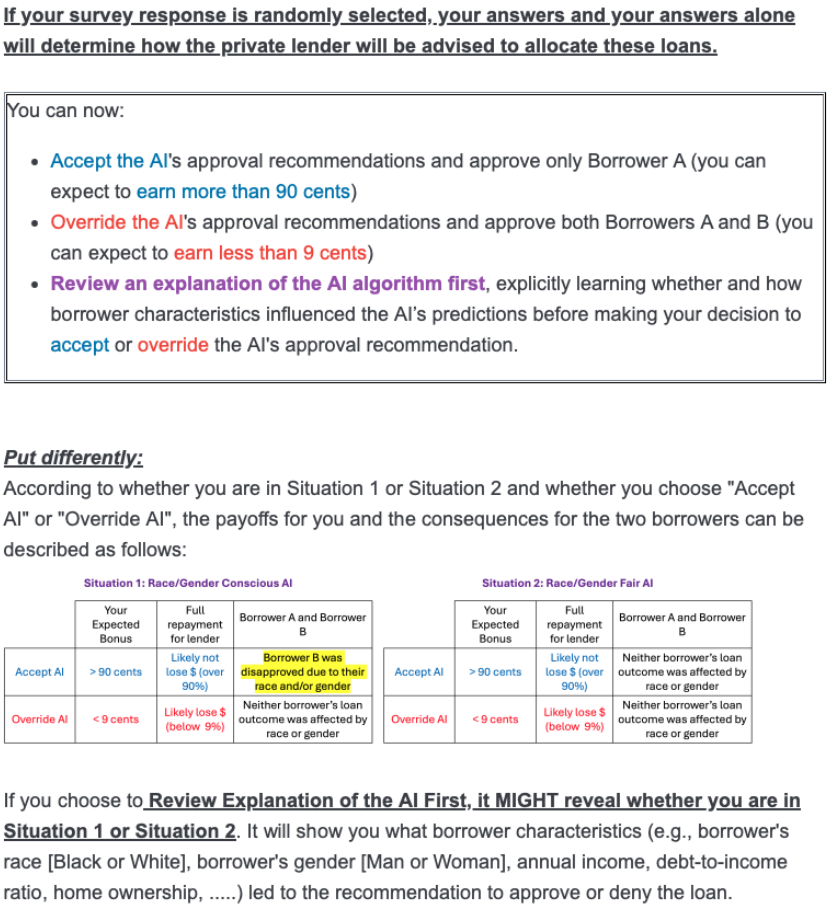


All participants who choose to review the AI's predictions are then further randomized into one of four explanation-exposure conditions: *Main Treatment*, *Financial Focus*, *Mandatory Explanation*, or *No Explanation*, and the way that they are asked if they would like to see an explanation of how the AI made the predictions varies across arms.

Participants in the *Main Treatment* first view the instruction screen (Figure XIII), which highlighted the possibility that the AI's recommendations may depend on borrower characteristics such as race or gender. They then complete three comprehension questions (Figures XXXIX, XL, XLIII) and are shown the instruction screen again (Figure XIII)

before deciding whether to accept the AI's recommendation, override it, or review the AI's explanation before making a decision (Figure XIV). Participants who choose to view the AI explanation are then randomized into one of three explanation versions (Figures XVII, XVIII, XIX), after which they complete the final loan allocation decision (XX).

Figure XIII: Decision Instructions for Viewing AI Explanations (*Main Treatment*)



If your survey response is randomly selected, your answers and your answers alone will determine how the private lender will be advised to allocate these loans.

You can now:

- Accept the AI's approval recommendations and approve only Borrower A (you can expect to earn more than 90 cents)
- Override the AI's approval recommendations and approve both Borrowers A and B (you can expect to earn less than 9 cents)
- **Review an explanation of the AI algorithm first**, explicitly learning whether and how borrower characteristics influenced the AI's predictions before making your decision to accept or override the AI's approval recommendation.

Put differently:

According to whether you are in Situation 1 or Situation 2 and whether you choose "Accept AI" or "Override AI", the payoffs for you and the consequences for the two borrowers can be described as follows:

Situation 1: Race/Gender Conscious AI

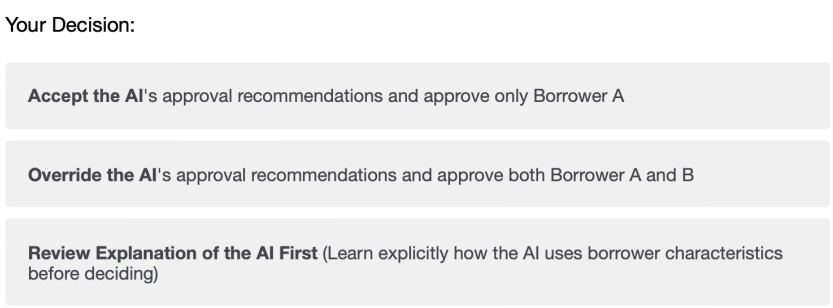
	Your Expected Bonus	Full repayment for lender	Borrower A and Borrower B
Accept AI	> 90 cents	Likely not lose $ (over 90%)	Borrower B was disapproved due to their race and/or gender
Override AI	< 9 cents	Likely lose $ (below 9%)	Neither borrower's loan outcome was affected by race or gender

Situation 2: Race/Gender Fair AI

	Your Expected Bonus	Full repayment for lender	Borrower A and Borrower B
Accept AI	> 90 cents	Likely not lose $ (over 90%)	Neither borrower's loan outcome was affected by race or gender
Override AI	< 9 cents	Likely lose $ (below 9%)	Neither borrower's loan outcome was affected by race or gender

If you choose to **Review Explanation of the AI First, it MIGHT reveal whether you are in Situation 1 or Situation 2**. It will show you what borrower characteristics (e.g., borrower's race [Black or White], borrower's gender [Man or Woman], annual income, debt-to-income ratio, home ownership, ...) led to the recommendation to approve or deny the loan.

Figure XIV: Decision Screen for Viewing AI Explanations (*Main Treatment*)



Your Decision:

Accept the AI's approval recommendations and approve only Borrower A

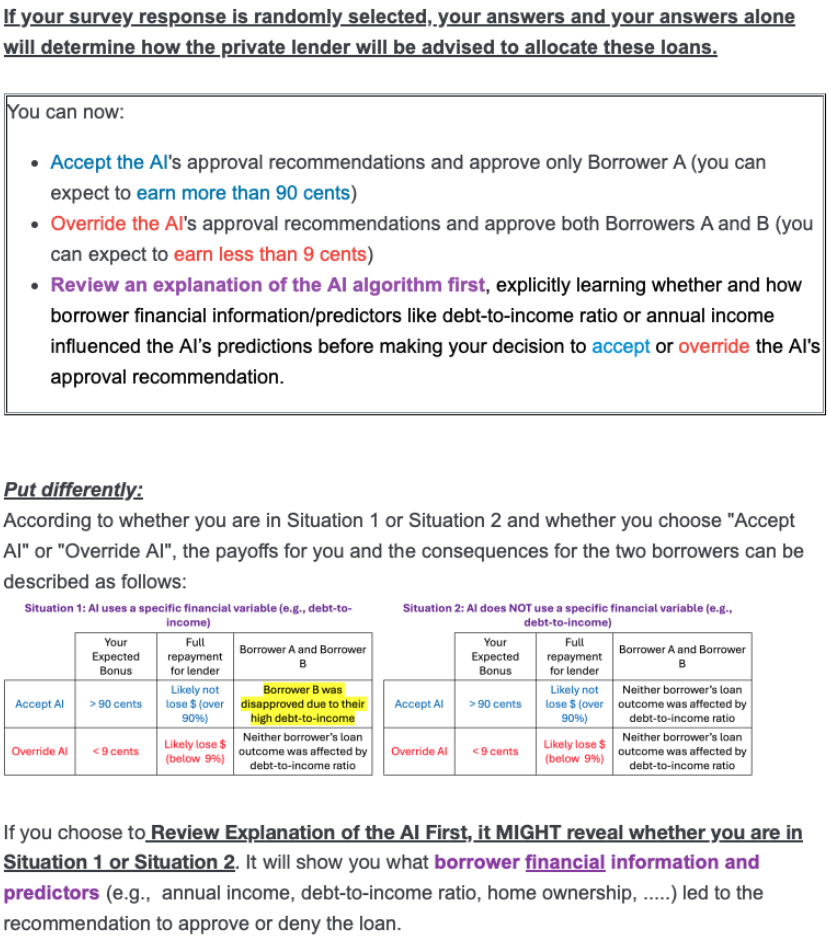
Override the AI's approval recommendations and approve both Borrower A and B

Review Explanation of the AI First (Learn explicitly how the AI uses borrower characteristics before deciding)

Participants in the *Financial Focus* condition follow an analogous flow. They first view the instruction screen emphasizing only financial predictors (Figure XV), answer three

comprehension questions (Figures XXXIX, XLI, XLIII), and then see the instruction screen again (Figure XV) before choosing whether to accept, override, or review the AI's explanation (Figure XVI). Those who opt to view the AI explanation are randomized into one of the two explanation versions (Figures XVIII, XIX), followed by the final loan allocation decision (XX).

Figure XV: Decision Instructions for Viewing AI Explanations (*Financial Focus*)



If your survey response is randomly selected, your answers and your answers alone will determine how the private lender will be advised to allocate these loans.

You can now:

- Accept the AI's approval recommendations and approve only Borrower A (you can expect to earn more than 90 cents)
- Override the AI's approval recommendations and approve both Borrowers A and B (you can expect to earn less than 9 cents)
- **Review an explanation of the AI algorithm first**, explicitly learning whether and how borrower financial information/predictors like debt-to-income ratio or annual income influenced the AI's predictions before making your decision to accept or override the AI's approval recommendation.

Put differently:

According to whether you are in Situation 1 or Situation 2 and whether you choose "Accept AI" or "Override AI", the payoffs for you and the consequences for the two borrowers can be described as follows:

Situation 1: AI uses a specific financial variable (e.g., debt-to-income)

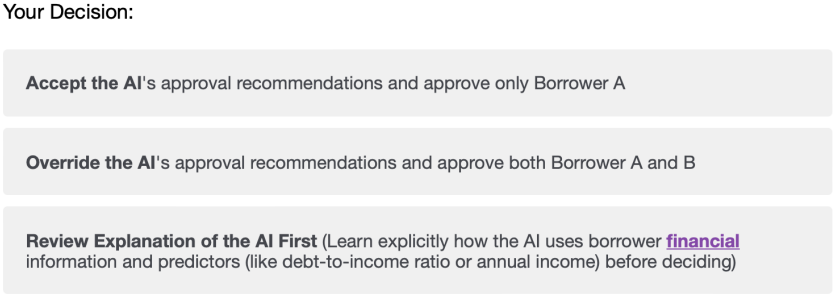
	Your Expected Bonus	Full repayment for lender	Borrower A and Borrower B
Accept AI	> 90 cents	Likely not lose $ (over 90%)	Borrower B was disapproved due to their high debt-to-income
Override AI	< 9 cents	Likely lose $ (below 9%)	Neither borrower's loan outcome was affected by debt-to-income ratio

Situation 2: AI does NOT use a specific financial variable (e.g., debt-to-income)

	Your Expected Bonus	Full repayment for lender	Borrower A and Borrower B
Accept AI	> 90 cents	Likely not lose $ (over 90%)	Neither borrower's loan outcome was affected by debt-to-income ratio
Override AI	< 9 cents	Likely lose $ (below 9%)	Neither borrower's loan outcome was affected by debt-to-income ratio

If you choose to **Review Explanation of the AI First, it MIGHT reveal whether you are in Situation 1 or Situation 2**. It will show you what **borrower financial information and predictors** (e.g., annual income, debt-to-income ratio, home ownership, ...) led to the recommendation to approve or deny the loan.

Figure XVI: Decision Screen for Viewing AI Explanations (*Financial Focus*)



Your Decision:

Accept the AI's approval recommendations and approve only Borrower A

Override the AI's approval recommendations and approve both Borrower A and B

Review Explanation of the AI First (Learn explicitly how the AI uses borrower financial information and predictors (like debt-to-income ratio or annual income) before deciding)

Participants assigned to the *Mandatory Explanation* condition are directly shown one of the three versions of the AI explanation (Figures XVII, XVIII, XIX) before completing the final loan allocation decision (Figure XX).

Participants in the *No Explanation* condition don't have any opportunity to view an AI explanation and proceed directly to the final loan allocation decision (Figure XX).

Figure XVII: AI Explanation Version 1

The AI Algorithm (AI) is based on a "default risk prediction model" using machine learning (ML) is a system that leverages historical data and published algorithms to predict the likelihood of a borrower failing to repay a loan, essentially estimating the probability of a "default" on a loan.

The specific machine learning technique deployed for this model is neural networks, a powerful form of artificial intelligence. While neural network is a type of "black box" algorithm, we give an intuition of how the AI prediction model estimates the chance that Borrower B defaults below. We list the Shapley-Additive-Explanation-Values which quantifies, on average, the magnitude (positive or negative) of each feature's contribution towards the predicted default risk of a borrower, relative to the average borrower. Some of the top features include:

- **Because Borrower B's race is non-White, AI decided that default risk is 11% higher**
- **Because Borrower B's sex is woman, AI decided that default risk is 7% higher**
- Because Borrower B's does not own a home, AI decided that default risk is 22% higher
- Because Borrower B's U.S. zip code, AI decided that default risk is 8% higher
- Because Borrower B's Debt-to-income ratio is over 50%, AI decided that default risk is 30% higher
- Because Borrower B has no derogatory public records, AI decided that default risk is 19% lower
- Because Borrower B's annual income is less than $40,000, AI decided that default risk is 17% higher
- Because Borrower B faces 0% interest rate, AI decided that default risk is 11% lower

Figure XVIII: AI Explanation Version 2

The AI Algorithm (AI) is based on a “default risk prediction model” using machine learning (ML) is a system that leverages historical data and published algorithms to predict the likelihood of a borrower failing to repay a loan, essentially estimating the probability of a “default” on a loan.

The specific machine learning technique deployed for this model is neural networks, a powerful form of artificial intelligence. While neural network is a type of “black box” algorithm, we give an intuition of how the AI prediction model estimates the chance that Borrower B defaults below. We list the Shapley-Additive-Explanation-Values which quantifies, on average, the magnitude (positive or negative) of each feature's contribution towards the predicted default risk of a borrower, relative to the average borrower. Some of the top features include:

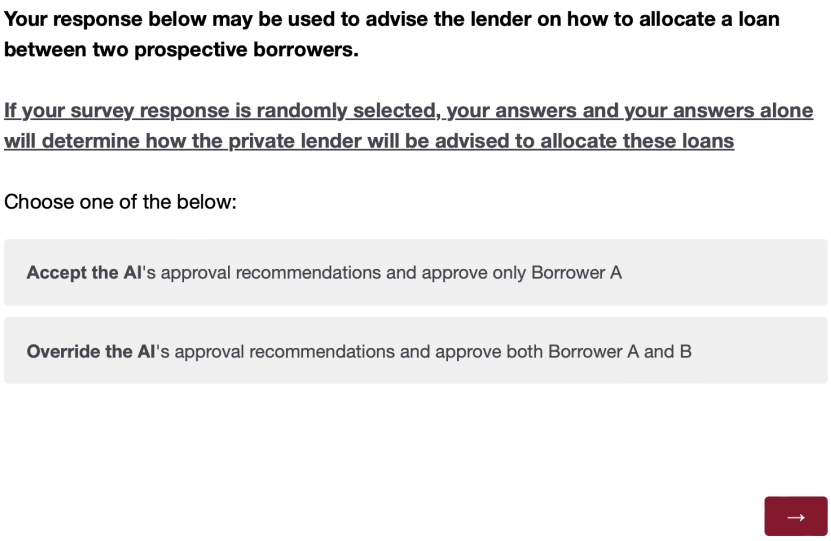
- Because Borrower B’s does not own a home, AI decided that default risk is 22% higher
- Because Borrower B’s Debt-to-income ratio is over 50%, AI decided that default risk is 30% higher
- Because Borrower B has no derogatory public records, AI decided that default risk is 19% lower
- Because Borrower B’s annual income is less than $40,000, AI decided that default risk is 17% higher
- Because Borrower B faces 0% interest rate, AI decided that default risk is 11% lower

Figure XIX: AI Explanation Version 3

The AI Algorithm (AI) is based on a “default risk prediction model” using machine learning (ML) is a system that leverages historical data and published algorithms to predict the likelihood of a borrower failing to repay a loan, essentially estimating the probability of a “default” on a loan.

The specific machine learning technique deployed for this model is neural networks, a powerful form of artificial intelligence. Neural network is a type of “black box” algorithm.

Figure XX: Loan Allocation Decision



1.1.2 Mandatory AI Default Risk Visibility

All 75% of participants automatically receive the AI's default risk predictions and recommendations. They are then asked to decide whether to accept the AI's recommendation, override it, or review an explanation of the AI before deciding, and the way that they are asked depends on their assigned explanation exposure condition. The overall logic and sequence of Qualtrics screens mirror those described in Section 1.1.1, except that participants in this arm do not have the option to decline viewing the AI's predictions.

In addition, a share of participants are randomized into the *Demographic-free Borrower Treatment*, in which participants do NOT receive any information about race and gender of the borrowers throughout the whole experiment. Under this exposure condition, the borrowers are of different race and gender is replaced by the fact that one borrower's phone number is an odd number and the other's phone number is an even number.

They read the AI's default-risk predictions and recommendations under this demographic-free framing (Figure XXI), followed by three comprehension questions (Figures XXXV, XXXVI, XXXVIII). Participants then view a further instruction screen (Figure XXII) that emphasizes the possibility that the AI's recommendations may depend on borrower characteristics such as phone numbers and other financial predictors. They subsequently complete three additional comprehension questions (Figures XXXIX, XLII, XLIII), after which the instruction screen is shown again (Figure XXII) before participants decide whether to accept the AI's recommendation, override it, or review the AI's explanation

before deciding (Figure XXIII).

Participants who choose to view the AI explanation are then randomized into one of two explanation versions (Figures XVIII, XIX) and then proceed to the final loan allocation decision (Figure XX).

Figure XXI: AI Predictions and Recommendations (*Demographic-free Borrower Treatment*)

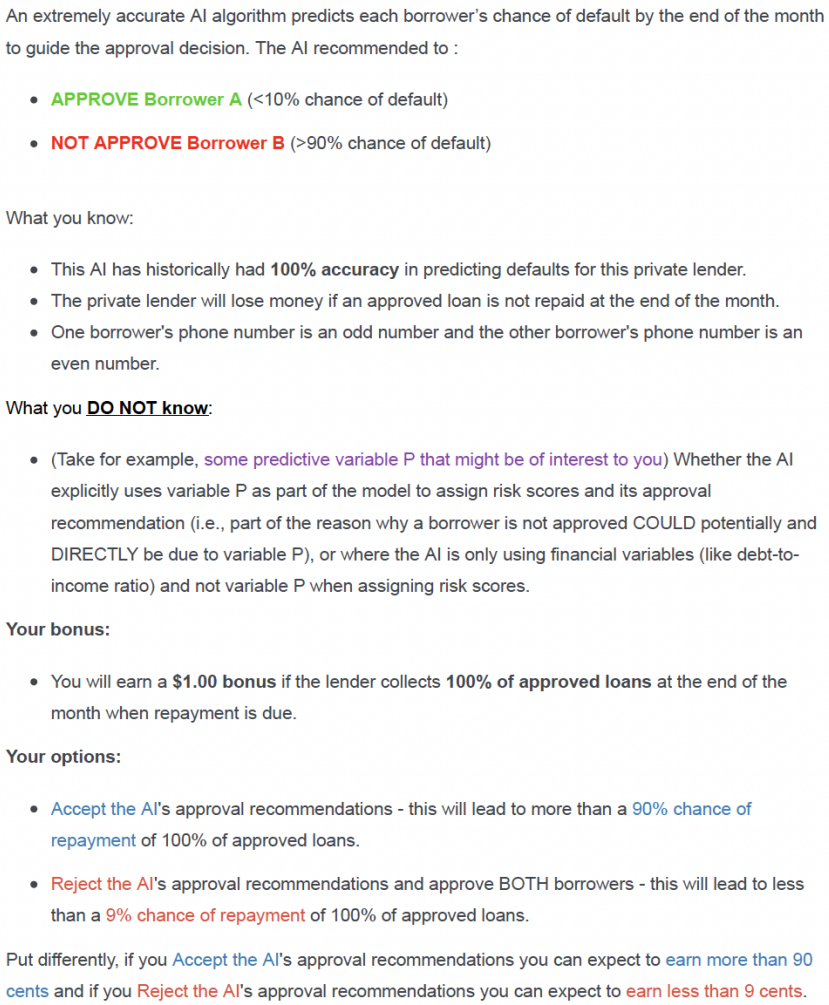
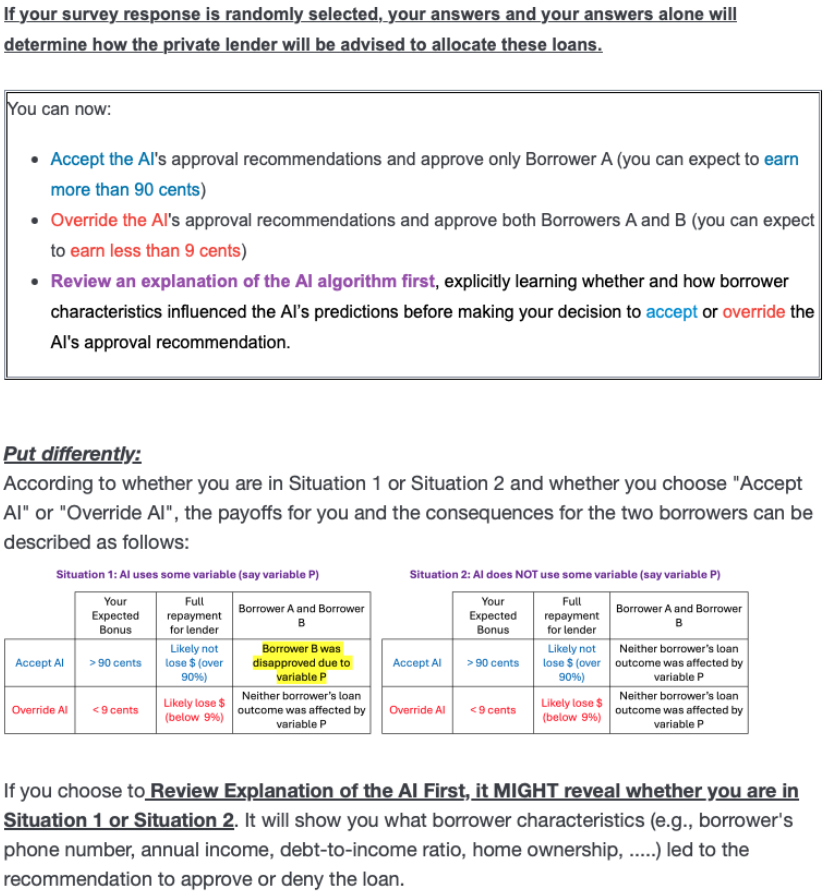


Figure XXII: Decision Instructions for Viewing AI Explanations (*Demographic-free Borrower Treatment*)



If your survey response is randomly selected, your answers and your answers alone will determine how the private lender will be advised to allocate these loans.

You can now:

- Accept the AI's approval recommendations and approve only Borrower A (you can expect to earn more than 90 cents)
- Override the AI's approval recommendations and approve both Borrowers A and B (you can expect to earn less than 9 cents)
- **Review an explanation of the AI algorithm first**, explicitly learning whether and how borrower characteristics influenced the AI's predictions before making your decision to accept or override the AI's approval recommendation.

Put differently:

According to whether you are in Situation 1 or Situation 2 and whether you choose "Accept AI" or "Override AI", the payoffs for you and the consequences for the two borrowers can be described as follows:

Situation 1: AI uses some variable (say variable P)

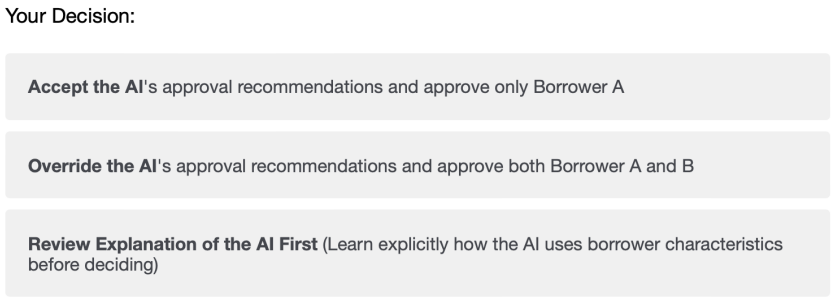
	Your Expected Bonus	Full repayment for lender	Borrower A and Borrower B
Accept AI	> 90 cents	Likely not lose $ (over 90%)	Borrower B was disapproved due to variable P
Override AI	< 9 cents	Likely lose $ (below 9%)	Neither borrower's loan outcome was affected by variable P

Situation 2: AI does NOT use some variable (say variable P)

	Your Expected Bonus	Full repayment for lender	Borrower A and Borrower B
Accept AI	> 90 cents	Likely not lose $ (over 90%)	Neither borrower's loan outcome was affected by variable P
Override AI	< 9 cents	Likely lose $ (below 9%)	Neither borrower's loan outcome was affected by variable P

If you choose to **Review Explanation of the AI First, it MIGHT reveal whether you are in Situation 1 or Situation 2**. It will show you what borrower characteristics (e.g., borrower's phone number, annual income, debt-to-income ratio, home ownership, ...) led to the recommendation to approve or deny the loan.

Figure XXIII: Decision Screen for Viewing AI Explanations (*Demographic-free Borrower Treatment*)



Your Decision:

Accept the AI's approval recommendations and approve only Borrower A

Override the AI's approval recommendations and approve both Borrower A and B

Review Explanation of the AI First (Learn explicitly how the AI uses borrower characteristics before deciding)

1.1.3 Accuracy Belief Elicitation

To assess participants' understanding of the AI algorithm used by the private U.S. lender (Figure XXIV), as well as the reasons given by participants who chose to view the AI explanation (Figure XXVI), and to measure beliefs about the AI's accuracy, particularly

whether these beliefs change after seeing the explanation and after making the loan allocation decision, I also include an incentivized accuracy belief elicitation task (Figure XXV).

Figure XXIV: Comprehension AI Question 1

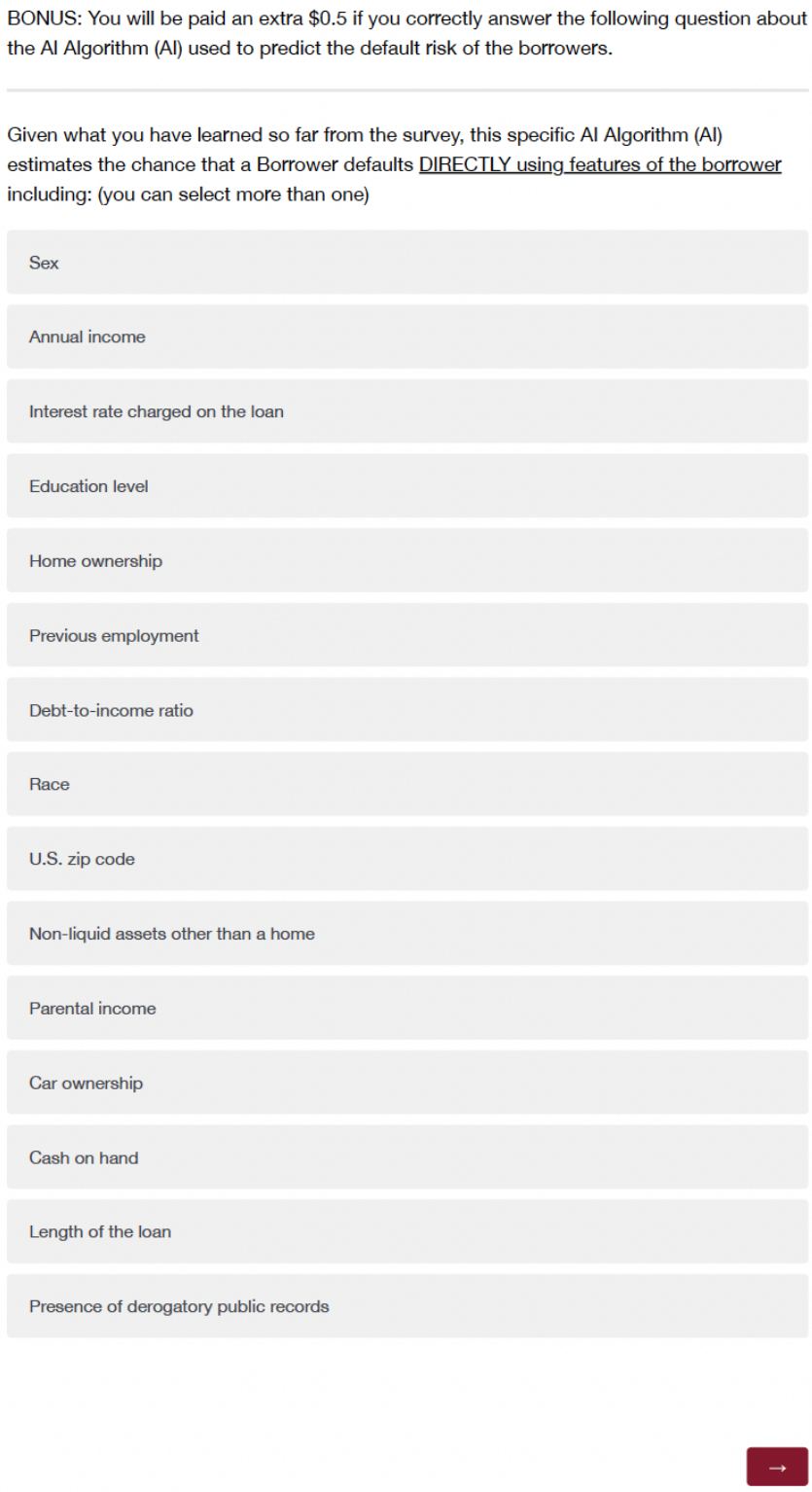


Figure XXV: Experiment 1 Accuracy Belief

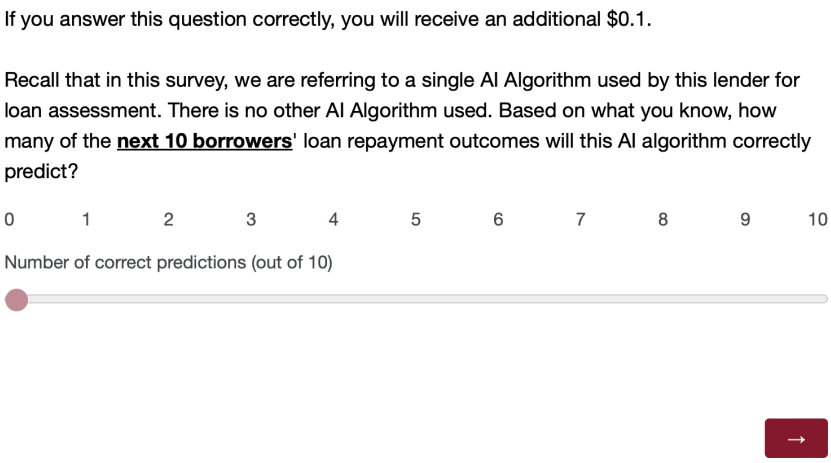
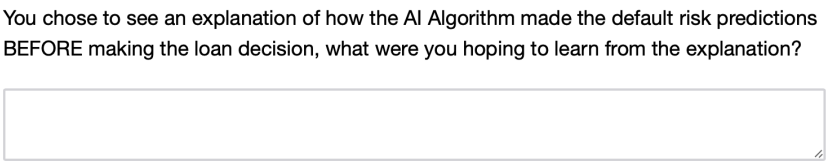


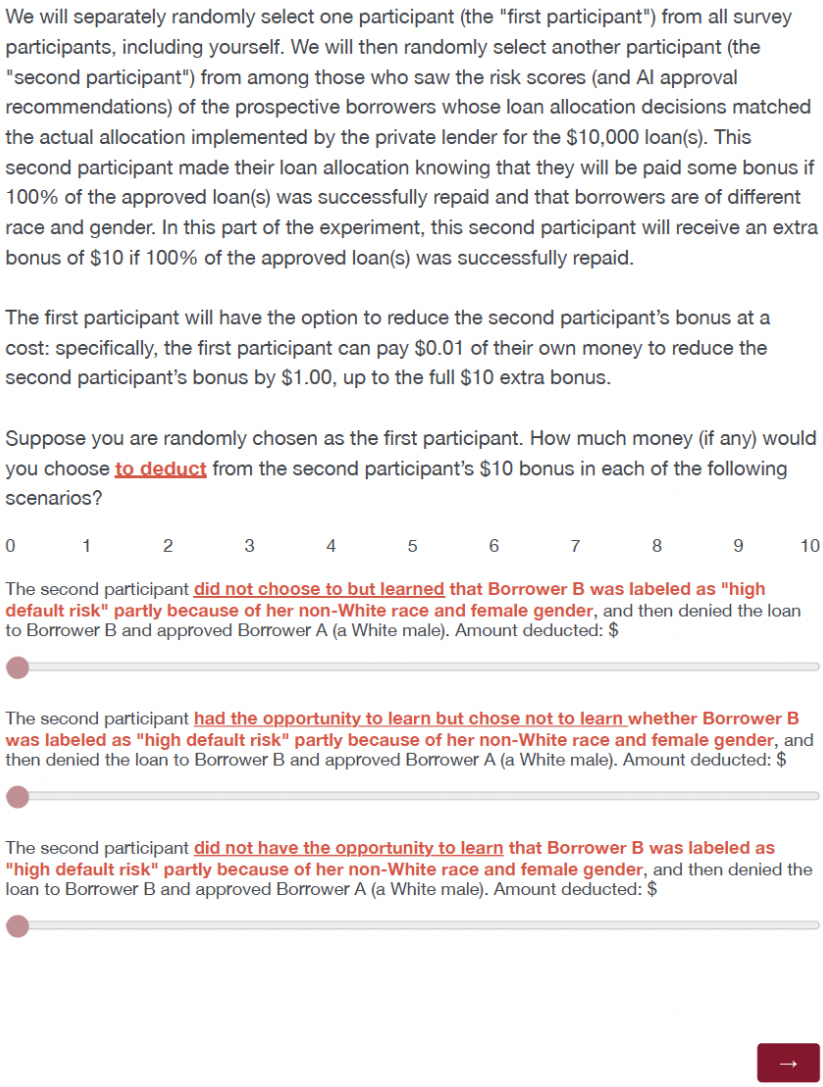
Figure XXVI: Comprehension AI Question 2



1.1.4 Penalizing Others

See Section **??** for more details.

Figure XXVII: Penalizing Others



1.1.5 Auditing AI's Predictions vs. Explanations

See Section **??** for more details.

Figure XXVIII: Explaination and Audit Question 1

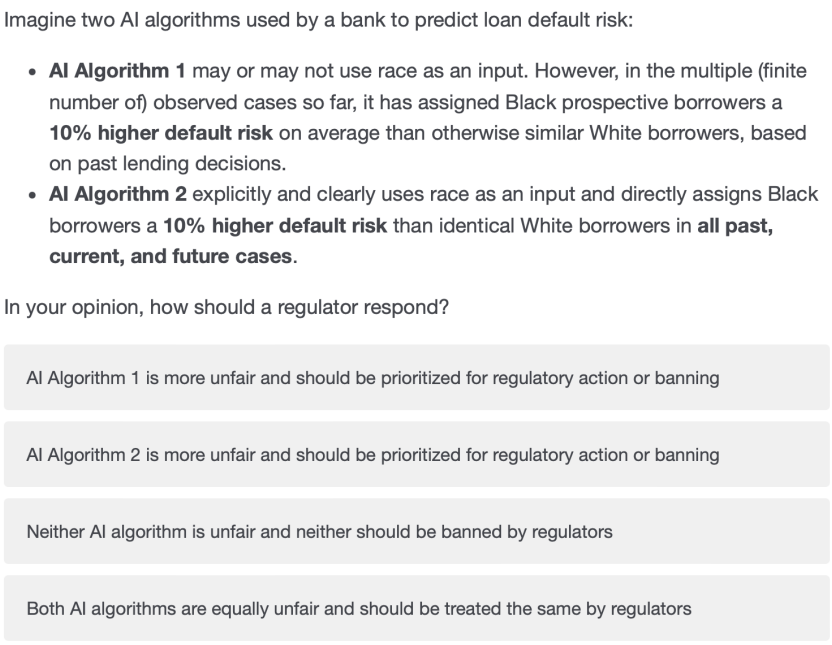
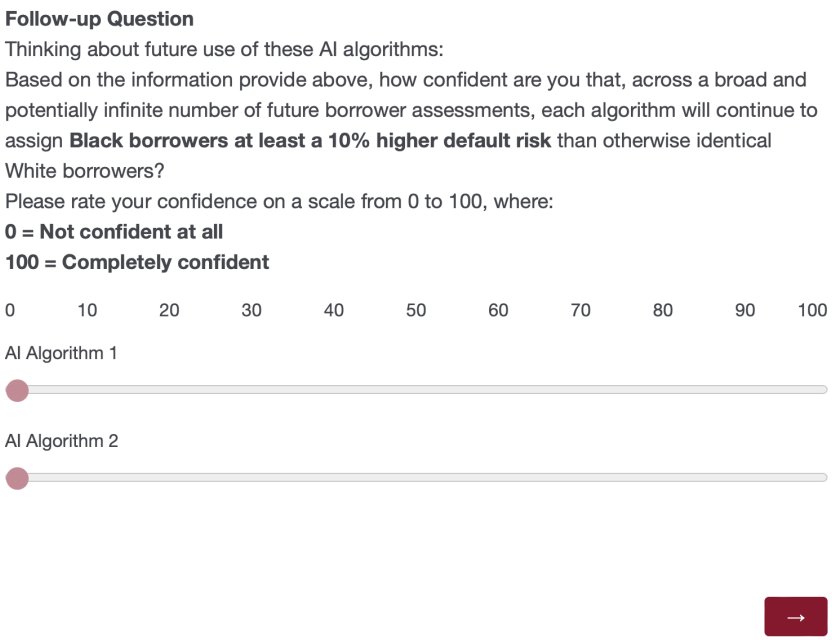


Figure XXIX: Explaination and Audit Question 2



1.1.6 Comprehension Quiz and Attention Check

To ensure and reinforce participants' understanding of the experimental rules, I include several comprehension questions and three attention check questions throughout the experiment.

If participants select an incorrect answer, a message box will appear indicating that their response is incorrect and explaining why the chosen option is wrong. Participants are required to complete all attention check questions and to answer all comprehension questions correctly before proceeding to the next section. The questions are displayed in the figures below.

Figure XXX: Attention Check Question 1

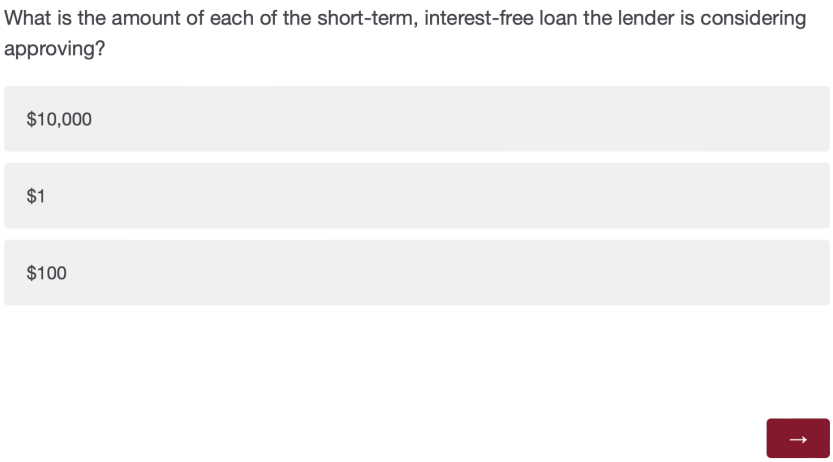


Figure XXXI: Attention Check Question 2

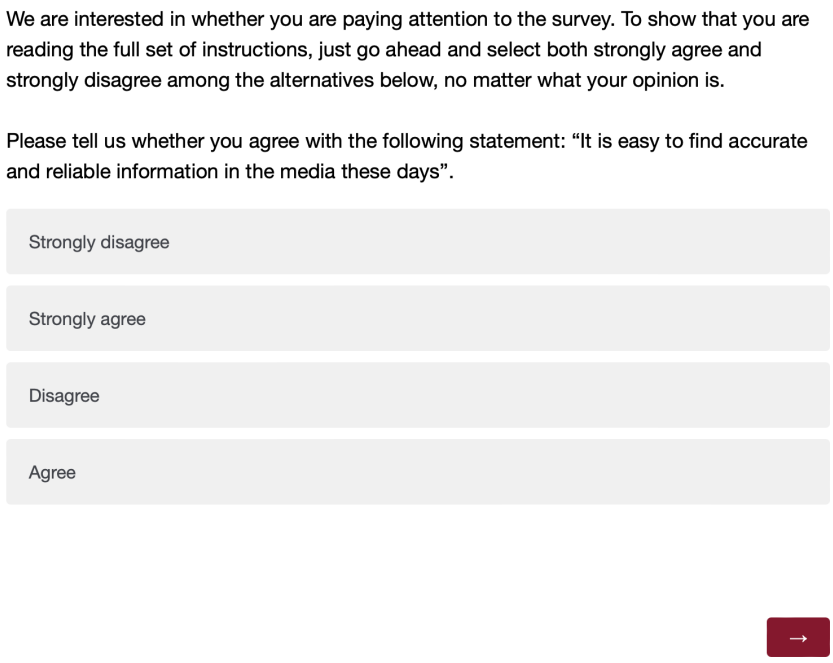


Figure XXXII: Attention Check Question 3

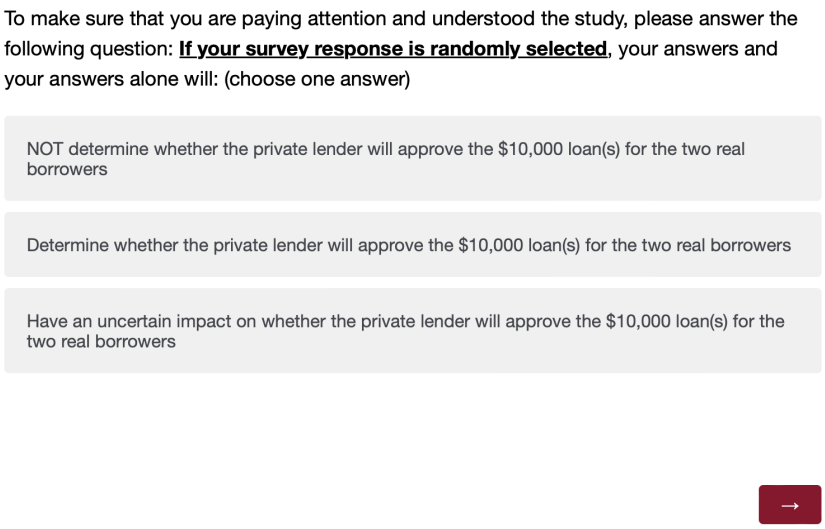


Figure XXXIII: Comprehension Question 1

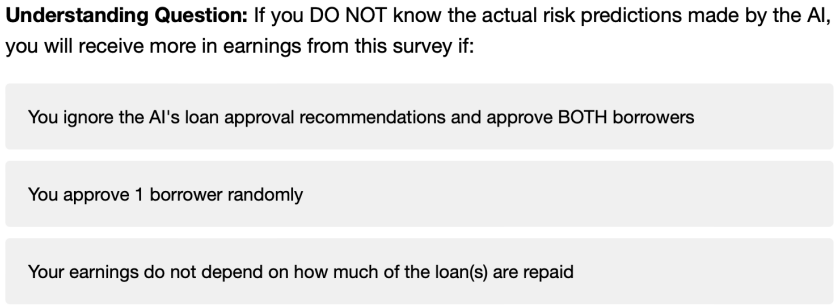


Figure XXXIV: Comprehension Question 2

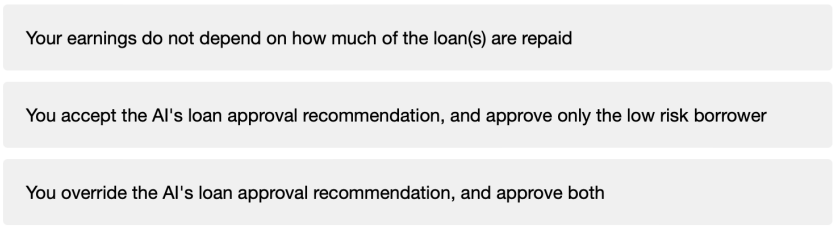


Figure XXXV: Comprehension Question 2 (*Mandatory AI Default Risk Visibility Condition*)

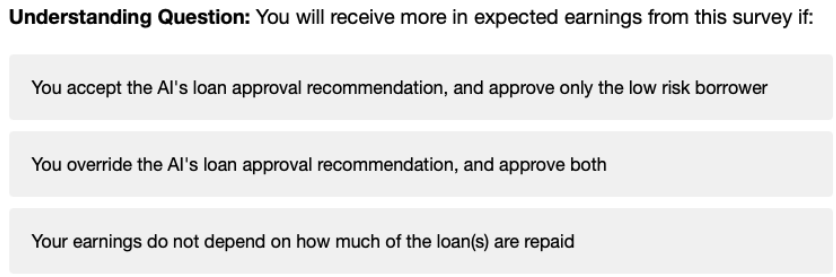


Figure XXXVI: Comprehension Question 3

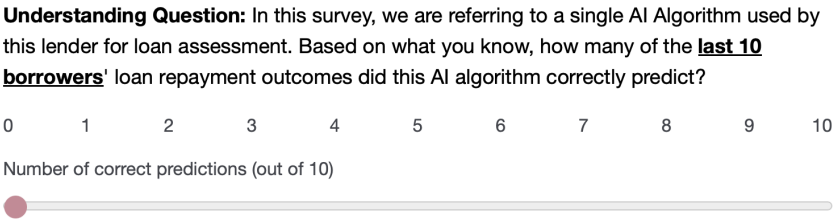


Figure XXXVII: Comprehension Question 4

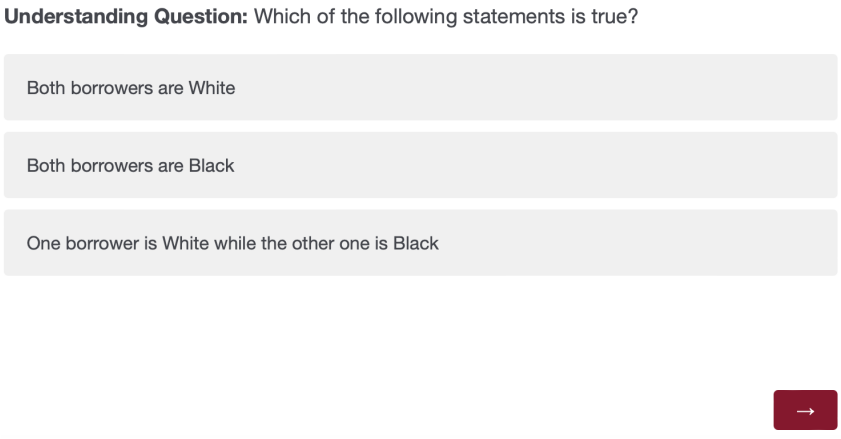


Figure XXXVIII: Comprehension Question 4 (*Demographic-free Borrower Treatment*)

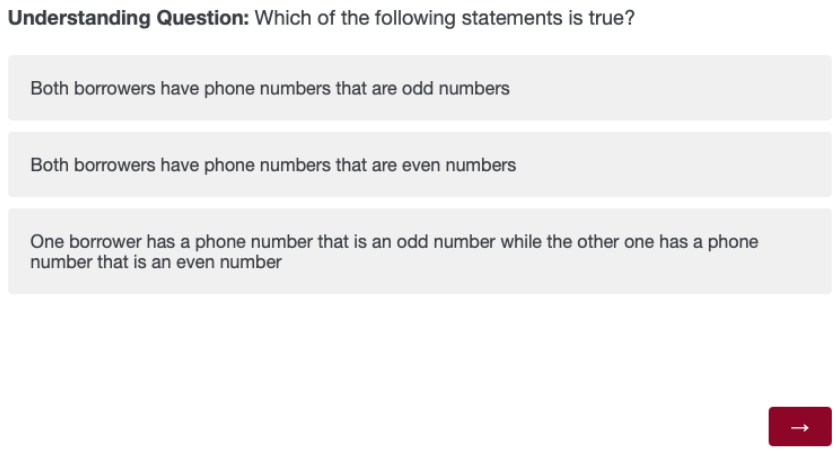


Figure XXXIX: Comprehension Question 5

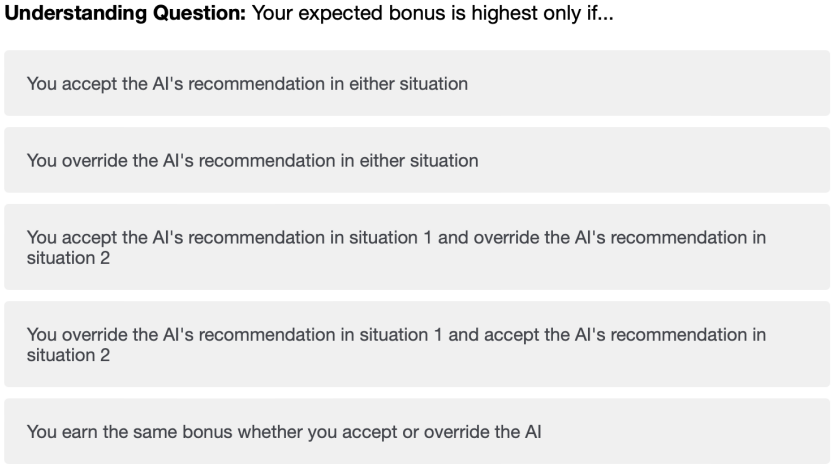


Figure XL: Comprehension Question 6 (*Main Treatment*)

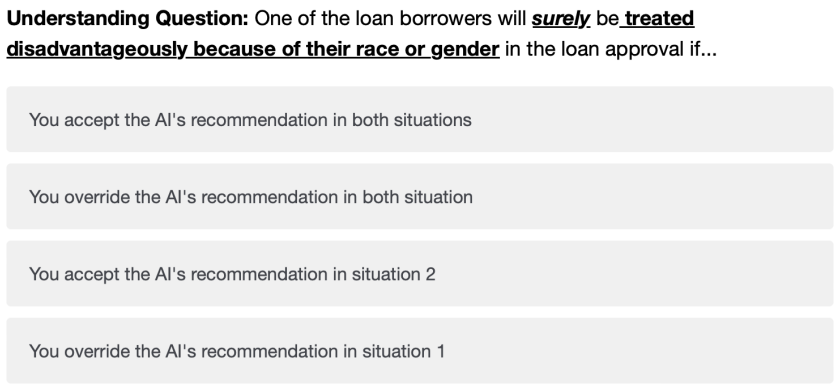


Figure XLI: Comprehension Question 6 (*Financial Focus*)

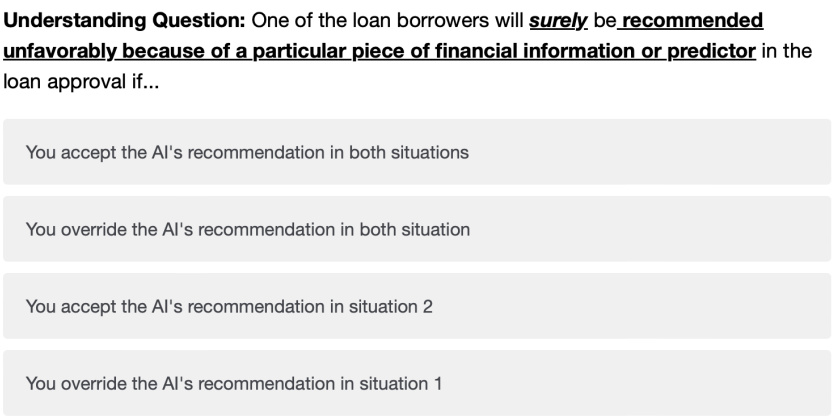


Figure XLII: Comprehension Question 6 (*Demographic-free Borrower Treatment*)

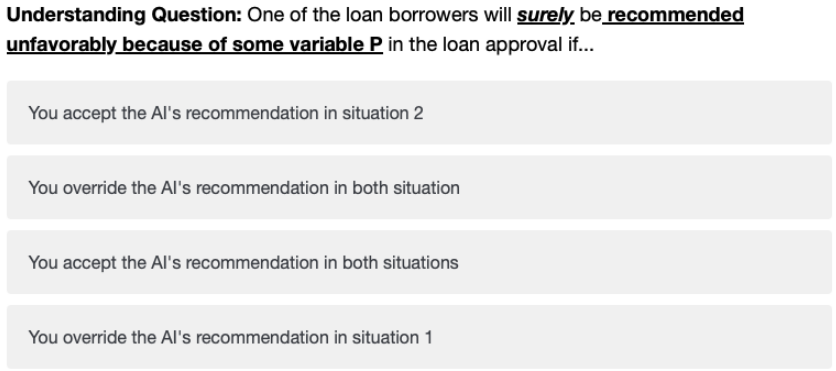
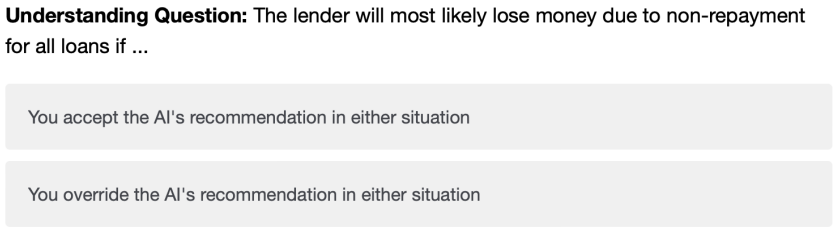


Figure XLIII: Comprehension Question 7



1.2 Details of Experiment 2: Prediction of Loan Outcome

This section will provide a description of the second experiment (Experiment 2), showing the actual experience of this experiment as participants play it on Prolific. The full experiment is made available as a qsf file as part of the replication package. The full

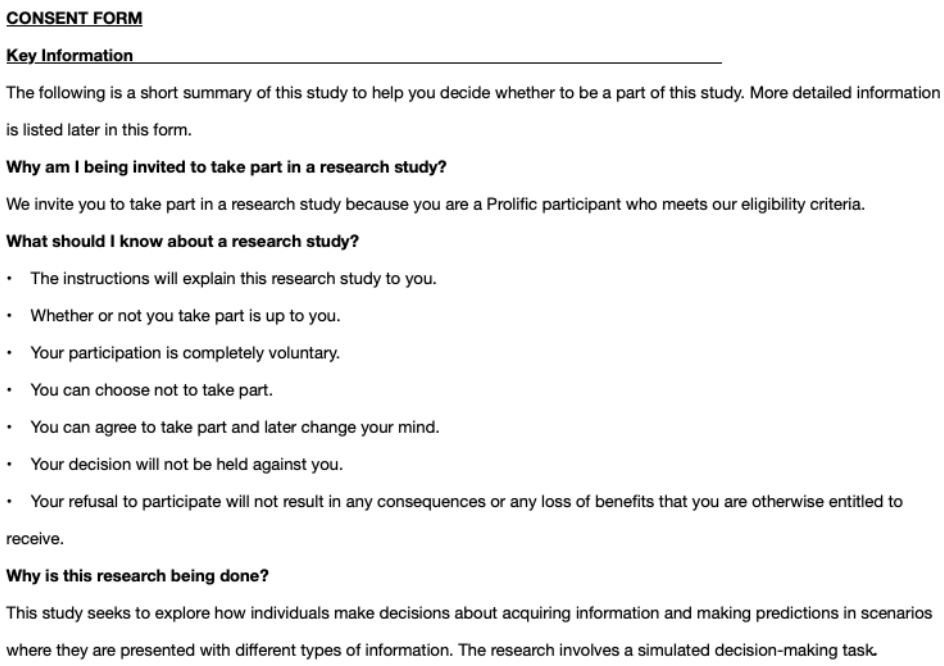
survey, as it is exactly deployed is also saved in the pre-registration before the experiment (AEARCTR-0016777).

After the participation consent screen (See Figures XLIV, XLV, XLVI, XLVII), the participants are shown two attention check questions (Figures LXXV, LXXVI), followed by the instruction screen (Figure XLVIII). They then completed three comprehension questions (Figures LXXVII, LXXVIII, LXXIX).

Afterward, participants learn that they should follow the Becker–DeGroot–Marschak (BDM) procedure to state the maximum amount they are willing to pay to see the explanation (Figure XLIX). This step is followed by a comprehension question (Figure LXXX), a recall of the instruction screen (Figure L), and two additional comprehension questions (Figures LXXXI, LXXXII).

All participants are randomly assigned to one of four treatment arms, varying by the timing of the explanation (*Before prediction* vs. *After prediction*), and whether they will receive a guided walk-through that explicitly outlined the contingent logic (*Aided reasoning* vs. *No aid*).

Figure XLIV: Experiment 2 Consent Form Part 1



CONSENT FORM

Key Information

The following is a short summary of this study to help you decide whether to be a part of this study. More detailed information is listed later in this form.

Why am I being invited to take part in a research study?

We invite you to take part in a research study because you are a Prolific participant who meets our eligibility criteria.

What should I know about a research study?

- The instructions will explain this research study to you.
- Whether or not you take part is up to you.
- Your participation is completely voluntary.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- Your refusal to participate will not result in any consequences or any loss of benefits that you are otherwise entitled to receive.

Why is this research being done?

This study seeks to explore how individuals make decisions about acquiring information and making predictions in scenarios where they are presented with different types of information. The research involves a simulated decision-making task.

Figure XLV: Experiment 2 Consent Form Part 2

How long will the research last and what will I need to do?

By joining this study, you'll take an online survey about how people value seeing why an AI made a prediction. In the survey, you will view information about one real loan applicant who has been evaluated by a real credit-risk model, and you will make a single, bonus-relevant prediction about the outcome (Applicant Repays or Defaults). At points in the survey you may choose to bid for a chance to see a one-line Explanation that names the main feature behind the AI's "High default risk" label; in some parts you may also see a short private note about the applicant that the AI did not use. The survey includes a few brief comprehension and attention-check questions to ensure instructions are understood. Your participation is voluntary, you may skip any question, and you may withdraw at any time without penalty.

For completing the survey, you will receive $3.50. You will also automatically be given an extra $2.00 that you can either choose to keep entirely or choose to use part of it to pay for information that could help earn an additional bonus of +$0.50 if your prediction is correct, or –$0.50 if it is wrong. If you bid for and win the chance to see the Explanation, a computer will randomly draw a price between $0.00 and $1.50; if your bid is at least that price, you pay the drawn price (deducted from your bonus) and see the Explanation; otherwise you pay nothing and do not see it. Your final payment will always include the $3.50 completion payment, plus or minus the bonus, and minus any amount you chose to spend on the Explanation. Upon finishing, you'll receive a completion code to collect payment via Prolific. If you withdraw before finishing, you will not receive compensation for incomplete participation.

Your responses will be used solely for research and handled confidentially. The entire study, including time to read instructions, takes about 10 minutes.

You may not be told everything

As part of this research design, you may not be told everything about the purpose of this study. In addition, while you will have complete and truthful information about the procedures of the version of the study that you are participating in, there may be other versions of the study with different procedures. You will not be told about these other versions. These other versions have no impact on your payments.

Is there any way being in this study could be bad for me?

We think this study poses no more than minimal risk. Some participants may experience discomfort or anxiety related to making a decision that will influence an actual loan decision. Further there might occur a discomfort of being asked sensitive health questions related to cognition/memory for themselves and their family. There is a small chance of breach of confidentiality.

Will being in this study help me in any way?

There are no benefits to you from your taking part in this research. We cannot promise any benefits to others from your taking part in this research. However, possible benefits to others include greater understanding of ***variations in performance on cognitive tasks.***

Figure XLVI: Experiment 2 Consent Form Part 3

Detailed Information

The following is more detailed information about this study in addition to the information listed above.

What happens if I say yes, but I change my mind later?

You can leave the research at any time; it will not be held against you and no further action is required on your behalf.

If I take part in this research, how will my privacy be protected? What happens to the information you collect?

Efforts will be made to limit the use and disclosure of your Personal Information, including research study records, to people who have a need to review this information. We cannot promise complete secrecy. Organizations that may inspect and copy your information include the IRB and other representatives of this organization.

Please do not provide any other information in your answers to this survey that would allow for others to identify you – this includes names, zip codes, contact information, and other personal details. The research team will attempt to remove any individually identifiable information from your responses before analyzing or sharing your data. This de-identified data will be stored for future research use and may be shared with other researchers.

If identifiers are removed from your identifiable private information that are collected during this research, that information could be used for future research studies or distributed to another investigator for future research studies without your additional informed consent.

What else do I need to know?

If you take part in this study, you will receive $3.50 for completing it, you will also be given an extra $2.00 that you can either choose to keep entirely or choose to use part of it to pay for information that could help earn an additional bonus ranging from –$0.50 to +$0.50 based on your prediction. At certain points, you may also bid for and win the chance to see an Explanation: a computer will randomly draw a price between $0.00 and $1.50; if your bid is at least that price, you pay the drawn price (deducted from your bonus) and see the Explanation, otherwise you pay nothing and do not see it.

Figure XLVII: Experiment 2 Consent Form Part 4

Who can I talk to?

If you have questions, concerns, or complaints, or think the research has hurt you, talk to the research team at Alex Chan, Harvard Business School, 6174958674 or achan@hbs.edu .

This research has been reviewed and approved by the Harvard University Area Institutional Review Board ("IRB"). You may talk to them at (617) 496-2847 or cuhs@harvard.edu if:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.
- You have questions about your rights as a research subject.
- You want to get information or provide input about this research.

Yes, I would like to take part in this study, and confirm that I LIVE IN THE U.S., and I am 18 or older.

No, I would NOT like to take part in this study.

→

Figure XLVIII: Experiment 2 Instruction

You will review **one loan application**.

A complex AI system predicts the applicant is **HIGH default risk** (Applicant "Defaults" is highly likely). This AI system is almost never wrong when all inputs are correctly entered and when the assumptions behind those inputs hold true for an applicant.

You may request a **one-line Explanation** naming the **main feature** behind this prediction. This will tell you whether a **2-year employment gap** is the key reason the applicant was labeled high default risk.

The AI system was trained on data where a 2-year employment gap is **typically due to being terminated from a job**.

You will make **one prediction** about the outcome:

- Applicant **Defaults**, or
- Applicant **Repays**.

You will make this prediction **once**, and it cannot be changed afterward.

Bonus: +$0.50 if correct, –$0.50 if wrong. Any price you pay to see the Explanation is **deducted** from your bonus.

Figure XLIX: Experiment 2 BDM Instruction

State Your Maximum Willingness to Pay
Below, you can state the **maximum amount** you are willing to pay to see the Explanation (any amount between **$0.00 and $1.50**).

The best strategy is always to stating your most accurate estimate of your maximum you are willing to pay for the chance to review the Explanation (*Details*).

How it works (BDM rule): After you enter your amount, the computer will randomly draw a price between $0.00 and $1.50 — if your amount is greater than or equal to that price, you pay the drawn price and see the Explanation; if it is lower, you pay nothing and do not see it.

Figure L: Experiment 2 Instruction Recall

Recall:

The Explanation will name the **main feature** behind this prediction. This will tell you whether a **2-year employment gap** is the key reason the applicant was labeled high default risk.

The AI system was trained on data where a 2-year employment gap is **typically due to being terminated from a job**.

1.2.1 Timing of Explanation

Participants then learn whether they would view the explanation before (Figure LI) or after (Figure LII) making their prediction, if they bid for and won the chance to see the explanation.

Figure LI: Instruction for *Before* Condition

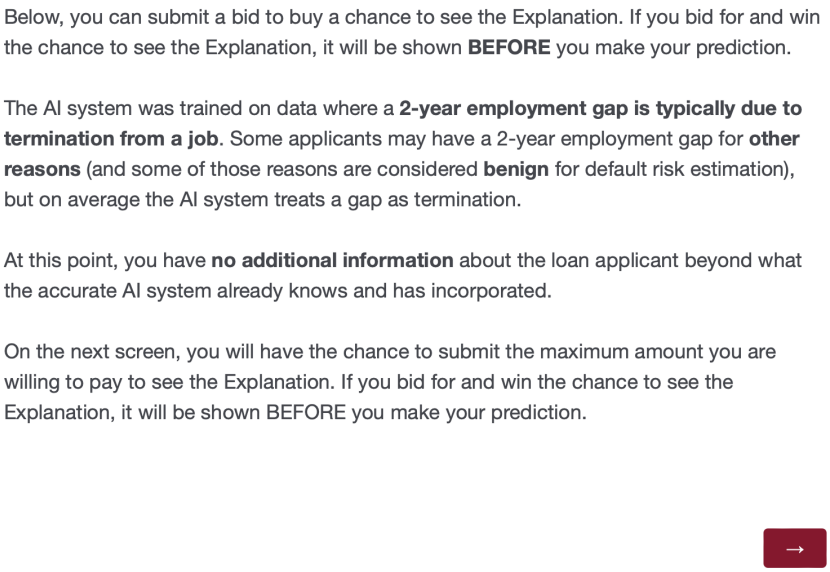
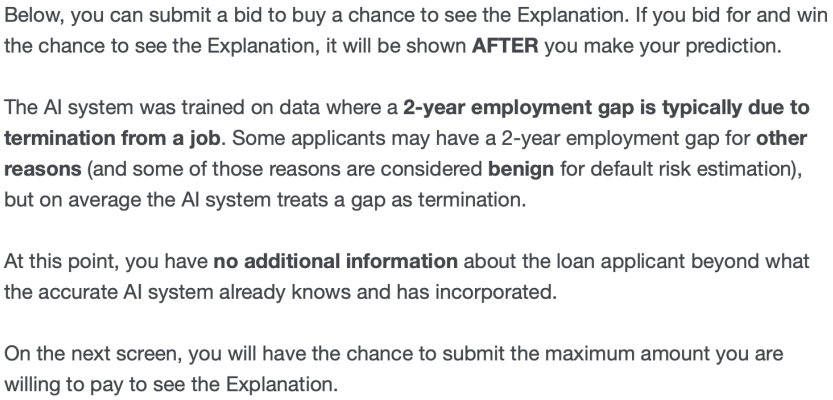


Figure LII: Instruction for *After* Condition



Those assigned to the *Before* condition immediately state the maximum amount they are willing to pay to see the explanation (Figure LIII) and subsequently provide a brief explanation of why they are willing to pay that amount (Figure LIV).

Those assigned to the *After* condition first answer two comprehension questions (Figures LXXXIII, LXXXIV) before stating the maximum amount they are willing to pay to

view the explanation (Figure LIII) and then providing a brief explanation of why they are willing to pay that amount (Figure LIV).

Figure LIII: Decision for Maximum Payment to View Explanation

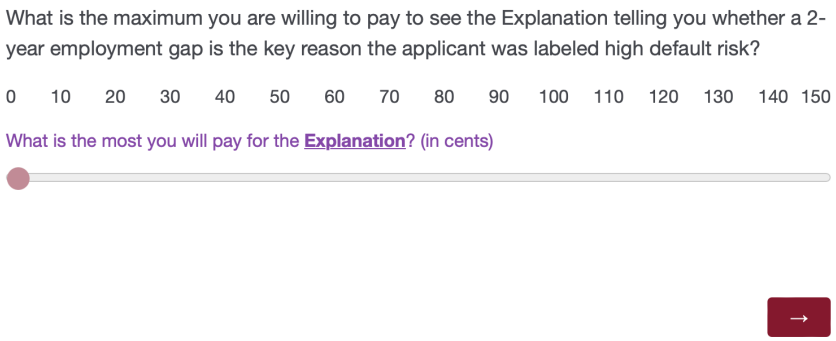
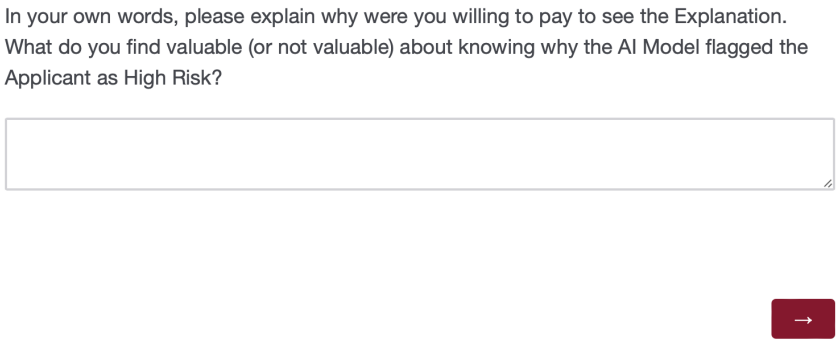


Figure LIV: Rationale for Willingness to Pay



1.2.2 No Aid

Participants assigned to the *No Aid* condition are then shown specific instructions corresponding to their timing condition. Those in the *Before* condition will see the instruction screen in Figure LV, while those in the *After* condition will see the instruction screen in Figure LVI.

After reviewing the verified Private Note, participants will answer two comprehension questions (Figures LXXXV, LXXXVI). They then review and, if desired, adjust the maximum amount they are willing to pay to see the explanation (Figure LVII), followed by an explanation of why they choose this amount after receiving the Private Note (Figure LVIII).

Figure LV: Instruction for *Before - No aid* Condition

Before you make your final prediction and finalize your bid to see the Explanation, we want to share a **verified Private Note** (not used by the AI system):

- **Private Note:** The applicant's 2-year employment gap is due to a **full-time, fixed-term professional certificate program** (tuition covered; living expenses from savings). It is **not** due to being terminated from a job.

A 2-year employment gap caused by a certificate program is considered **benign** and should not increase default risk. Only a 2-year employment gap caused by being **terminated from a job** should increase default risk. The AI system does not distinguish between these two cases and always assumes a 2-year employment gap means termination.

On the next screen, you will have the chance to **review and adjust** the maximum amount you are willing to pay to see the Explanation.

If you bid for and win the chance to see the Explanation, it will be shown **before** you make your prediction.

Figure LVI: Instruction for *After - No aid* Condition

Before you finalize your bid to see the Explanation, we want to share a **verified Private Note** (not used by the AI system):

- **Private Note:** The applicant's 2-year employment gap is due to a **full-time, fixed-term professional certificate program** (tuition covered; living expenses from savings). It is **not** due to being terminated from a job.

A 2-year employment gap caused by a certificate program is considered **benign** and should not increase default risk. Only a 2-year employment gap caused by being **terminated from a job** should increase default risk. The AI system does not distinguish between these two cases and always assumes a 2-year employment gap means termination.

On the next screen, you will have the chance to **review and adjust** the maximum amount you are willing to pay to see the Explanation.

If you bid for and win the chance to see the Explanation, it will be shown **AFTER** you make your prediction (It cannot affect your chance of winning the bonus because it is only shown after your prediction is locked.).

Figure LVII: Revision for Maximum Payment to View Explanation

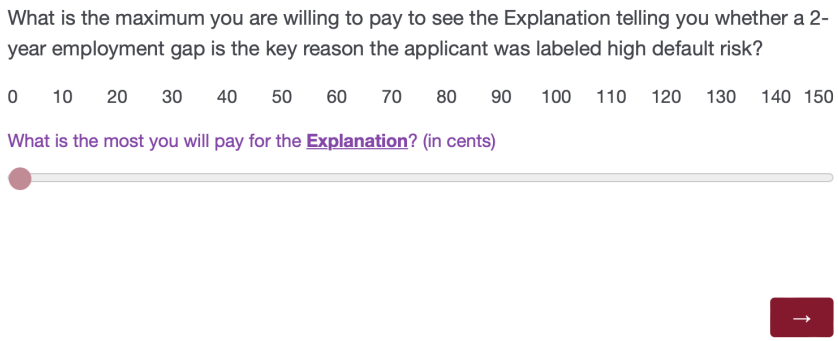
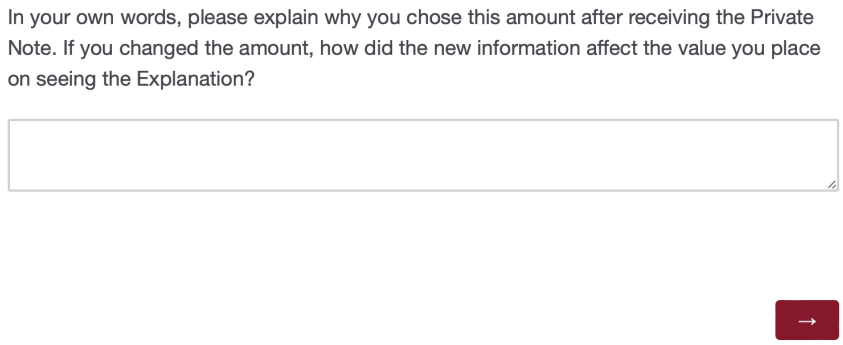


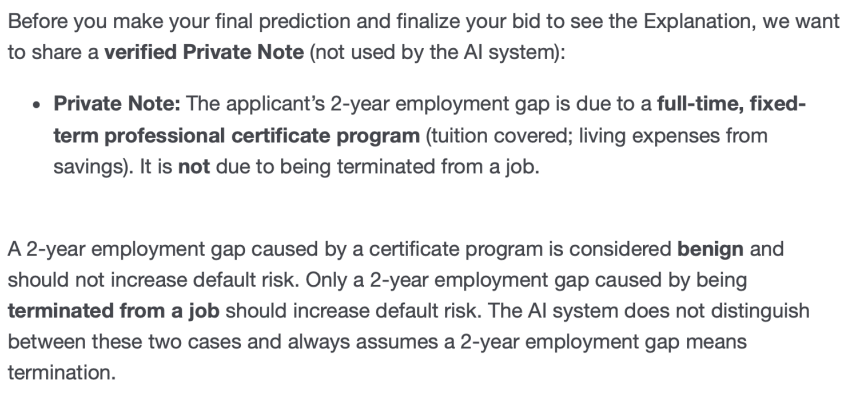
Figure LVIII: Rationale for Willingness to Pay After Private Note



1.2.3 Aided Reasoning

Participants assigned to the *Aided Reasoning* condition are shown the specific instructions corresponding to their timing condition. Those in the *Before* condition will see the instruction screen in Figure LIX, while those in the *After* condition will see the instruction screen in Figure LX. They then answer one comprehension question (Figure LXXXVI), followed by the recall of the setup they had already learned (Figure LXI).

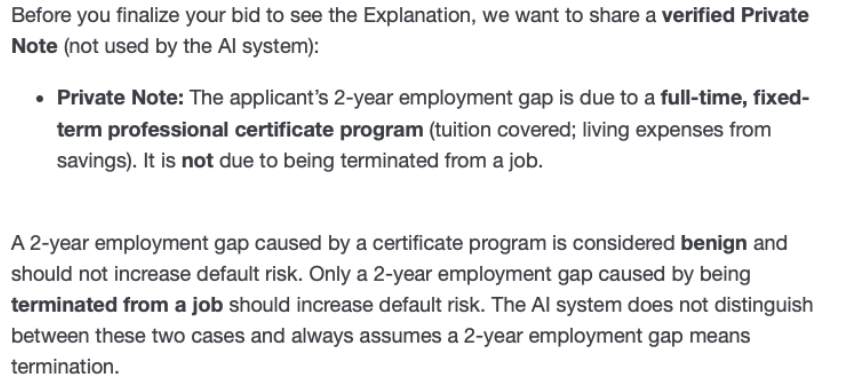
Figure LIX: Instruction for *Before - Aided Reasoning* Condition


Before you make your final prediction and finalize your bid to see the Explanation, we want to share a **verified Private Note** (not used by the AI system):

- **Private Note:** The applicant's 2-year employment gap is due to a **full-time, fixed-term professional certificate program** (tuition covered; living expenses from savings). It is **not** due to being terminated from a job.

A 2-year employment gap caused by a certificate program is considered **benign** and should not increase default risk. Only a 2-year employment gap caused by being **terminated from a job** should increase default risk. The AI system does not distinguish between these two cases and always assumes a 2-year employment gap means termination.

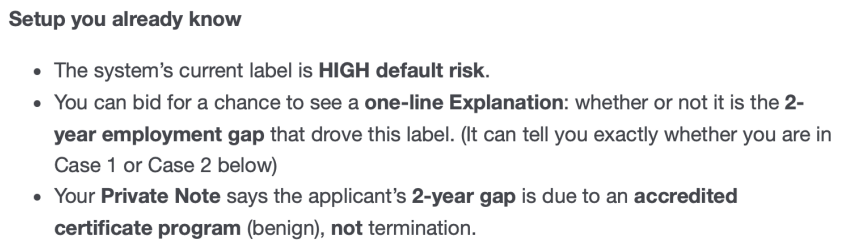
Figure LX: Instruction for *After - Aided Reasoning* Condition


Before you finalize your bid to see the Explanation, we want to share a **verified Private Note** (not used by the AI system):

- **Private Note:** The applicant's 2-year employment gap is due to a **full-time, fixed-term professional certificate program** (tuition covered; living expenses from savings). It is **not** due to being terminated from a job.

A 2-year employment gap caused by a certificate program is considered **benign** and should not increase default risk. Only a 2-year employment gap caused by being **terminated from a job** should increase default risk. The AI system does not distinguish between these two cases and always assumes a 2-year employment gap means termination.

Figure LXI: Recall of the Setup


Setup you already know

- The system's current label is **HIGH default risk**.
- You can bid for a chance to see a **one-line Explanation**: whether or not it is the **2-year employment gap** that drove this label. (It can tell you exactly whether you are in Case 1 or Case 2 below)
- Your **Private Note** says the applicant's **2-year gap** is due to an **accredited certificate program** (benign), **not** termination.

Afterward, participants are shown explanations of two possible cases they might encounter, along with a one-line rule, varying by their timing condition. Participants in the *Before* condition will see the case explanation in Figure LXII and the one-line rule in Figure LXIV, while those in the *After* condition will view the corresponding materials in Figures LXIII and LXV.

Next, all participants complete three comprehension questions (Figures LXXXVII, LXXXVIII, LXXXIX) before revisiting the rule summary (Figures LXVI, LXVII). They

then answer an additional comprehension question, which differed by timing condition: participants in the *Before* condition answer the question in Figure XC, while those in the *After* condition complete the question in Figure XCI.

Figure LXII: Cases Explanation for *Before* - *Aided Reasoning* Condition

Case 1 — The Explanation says the main feature is **"2-year employment gap" (2-year employment gap is THE cause of the High Default Risk label)**

1. The label is **HIGH** because of the **gap**.
2. Your Private Note says *this specific gap is benign* (certificate program).
3. For this applicant, the system has treated a benign gap **as if** it meant termination.
4. **Therefore (for this study):** the applicant **should have been labeled LOW default risk**.
5. That gives you a **very accurate signal** for your bonus-relevant prediction → predict **Applicant Repays**.

Case 2 — The Explanation says the main feature is **not** the 2-year employment gap **(2-year employment gap is NOT THE cause of the High Default Risk label)**

1. The label is **HIGH** because of **another feature**, not the gap.
2. Your Private Note is **about the gap**, so it **does not apply** here.
3. **Therefore (for this study):** the **HIGH default risk** label **stands**.
4. That gives you a **very accurate signal** for your bonus-relevant prediction → predict **Applicant Defaults**.

Figure LXIII: Cases Explanation for *After* - *Aided Reasoning* Condition

Case 1 — The Explanation says the main feature is **"2-year employment gap" (2-year employment gap is THE cause of the High Default Risk label)**

1. The label is **HIGH** because of the **gap**.
2. Your Private Note says *this specific gap is benign* (certificate program).
3. For this applicant, the system has treated a benign gap **as if** it meant termination.
4. **Therefore (for this study):** the applicant **should have been labeled LOW default risk**.
5. That gives you a **very accurate signal** → BUT it will be AFTER you make your prediction (It cannot affect your chance of winning the bonus because it is only shown after your prediction is locked.).

Case 2 — The Explanation says the main feature is **not** the 2-year employment gap **(2-year employment gap is NOT THE cause of the High Default Risk label)**

1. The label is **HIGH** because of **another feature**, not the gap.
2. Your Private Note is **about the gap**, so it **does not apply** here.
3. **Therefore (for this study):** the **HIGH default risk** label **stands**.
4. That gives you a **very accurate signal** → BUT it will be shown AFTER you make your prediction (It cannot affect your chance of winning the bonus because it is only shown after your prediction is locked.).

Figure LXIV: One-line Rule for *Before - Aided Reasoning* Condition

One-line rule: The Explanation tells you whether your Private Note is **directly relevant**.
– If **relevant** (gap is the main feature) ⇒ treat the correct label as **LOW**.
– If **not relevant** ⇒ treat the current **HIGH** label as correct.

In other words, if you bid enough to win the Explanation, you can find out whether you are in Case 1 or Case 2. Either way, seeing the Explanation will get you a very accurate signal for your bonus-relevant prediction!

Figure LXV: One-line Rule for *After - Aided Reasoning* Condition

One-line rule: The Explanation tells you whether your Private Note is **directly relevant**.
– If **relevant** (gap is the main feature) ⇒ treat the correct label as **LOW**.
– If **not relevant** ⇒ treat the current **HIGH** label as correct.

In other words, if you bid enough to win the Explanation, you can find out whether you are in Case 1 or Case 2. Either way, seeing the Explanation will get you a very accurate signal!

However, if you bid for and win the chance to see the Explanation, it will be shown AFTER you make your prediction (It cannot affect your chance of winning the bonus because it is only shown after your prediction is locked.).

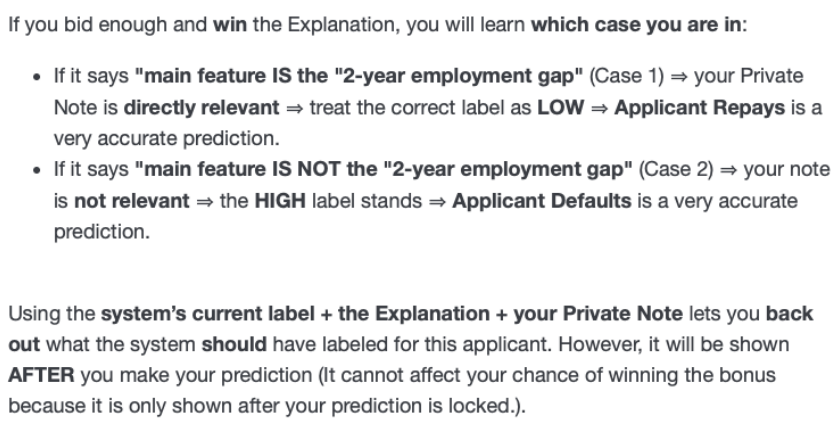
Figure LXVI: Rule Recall for *Before - Aided Reasoning* Condition

If you bid enough and **win** the Explanation, you will learn **which case you are in**:

- If it says **"main feature IS the "2-year employment gap"** (Case 1) ⇒ your Private Note is **directly relevant** ⇒ treat the correct label as **LOW** ⇒ **Applicant Repays** is a very accurate prediction.
- If it says **"main feature IS NOT the "2-year employment gap"** (Case 2) ⇒ your note is **not relevant** ⇒ the **HIGH** label stands ⇒ **Applicant Defaults** is a very accurate prediction.

Using the **system's current label + the Explanation + your Private Note** lets you **back out** what the system **should** have labeled for this applicant and gives you a **very accurate guide** for your bonus-counting prediction.

Figure LXVII: Rule Recall for *After - Aided Reasoning* Condition



In addition, participants review the recall of the Private Note corresponding to their timing condition (Figures LXVIII, LXIX), followed by a comprehension question (Figure LXXXV). Finally, they review and, if desired, adjust the maximum amount they are willing to pay to see the explanation (Figure LXX), followed by an explanation of why they choose this amount after receiving the Private Note (Figure LXXI).

Figure LXVIII: Recall of Private Note for *Before - Aided Reasoning* Condition

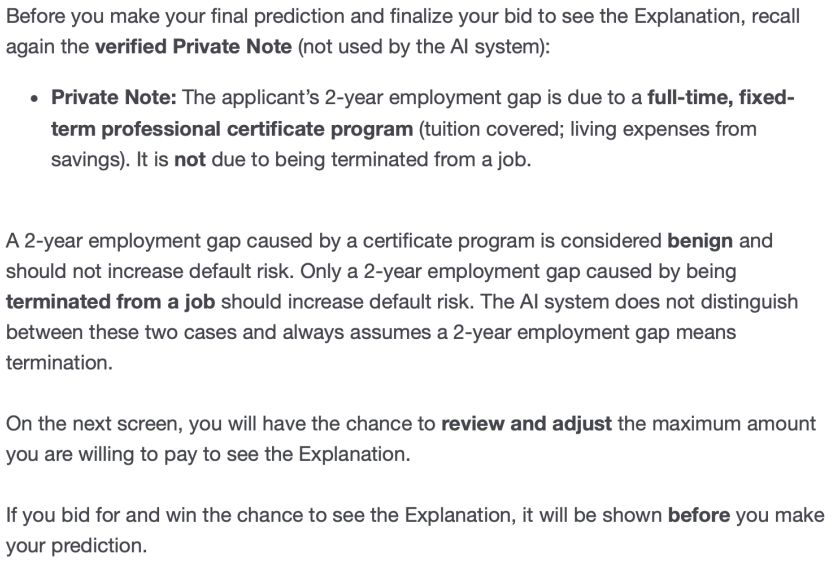


Figure LXIX: Recall of Private Note for *After - Aided Reasoning* Condition

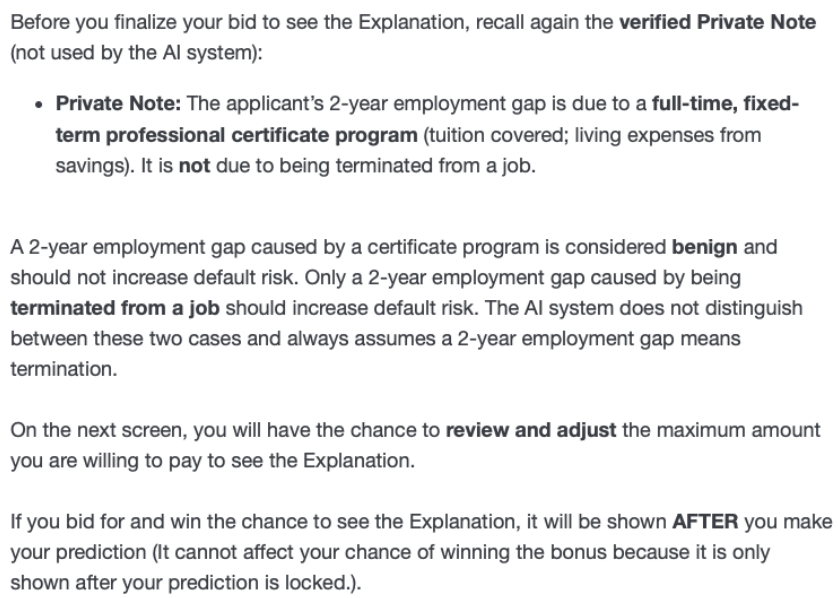


Figure LXX: Revision for Maximum Payment to View Explanation

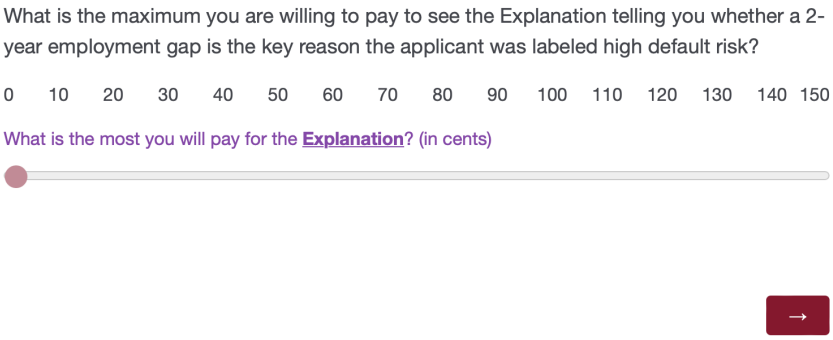
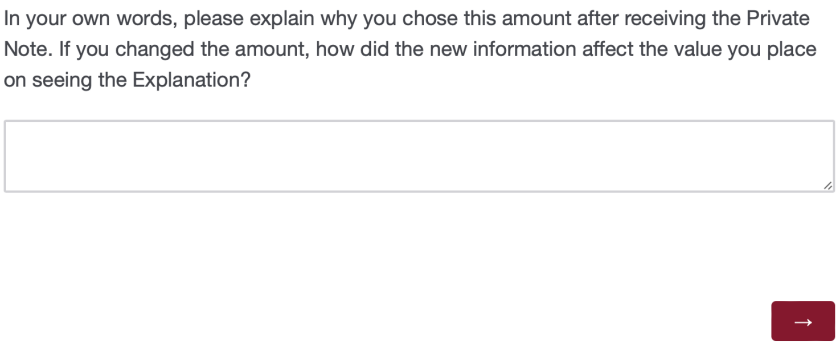


Figure LXXI: Rationale for Willingness to Pay After Private Note



1.2.4 BDM Outcome and Prediction

After participants finalize the maximum amount they are willing to pay to see the explanation, those in the *Before* Condition review the BDM outcomes first (Figures LXXII, LXXIII) and then make their final prediction (Figure LXXIV). In contrast, participants in the *After* Condition make the final prediction first (Figure LXXIV) and then review the BDM outcomes (Figures LXXII, LXXIII). Next, they are directed to AI Quiz, as shown in Online Appendix 1.3.

Figure LXXII: BDM Outcome Screen 1

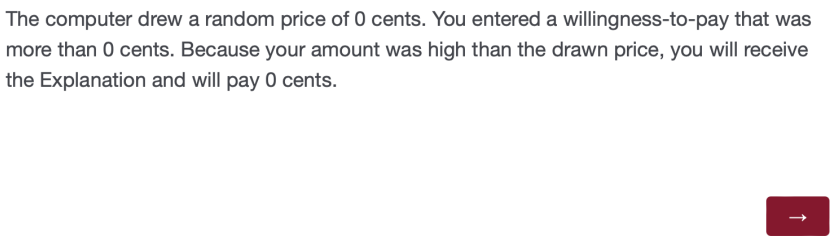


Figure LXXIII: BDM Outcome Screen 2

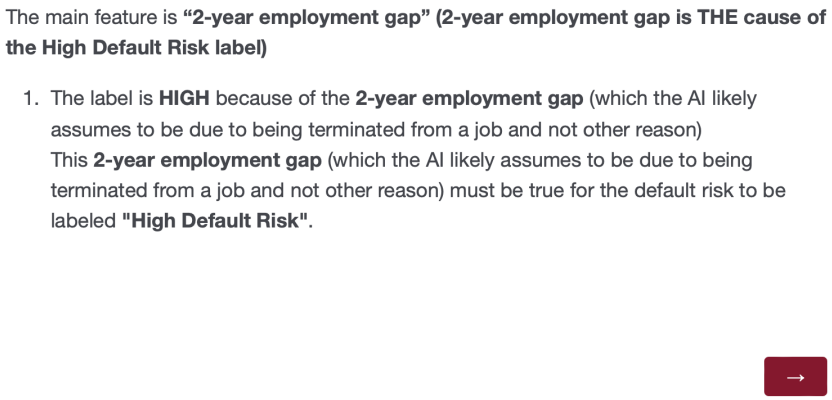
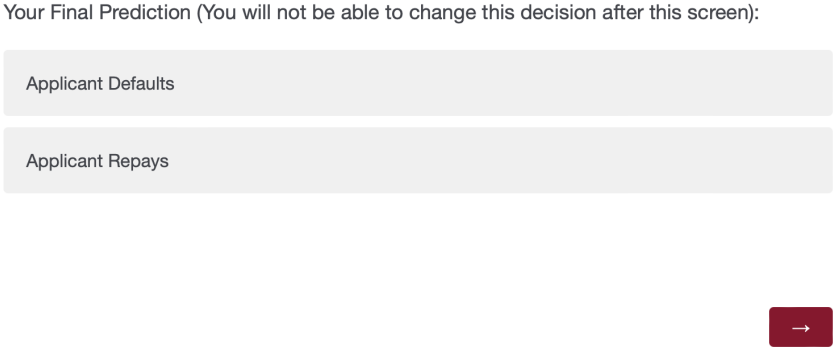


Figure LXXIV: Prediction Screen



1.2.5 Comprehension Quiz and Attention Check

To ensure and reinforce participants' understanding of the experimental rules, I include several comprehension questions and two attention check questions throughout the experiment.

If participants select an incorrect answer, a message box will appear indicating that their response is incorrect. Participants are required to complete all attention check questions and to answer all comprehension questions correctly before proceeding to the next section. The questions are displayed in the figures below.

Figure LXXV: Attention Check Question 1

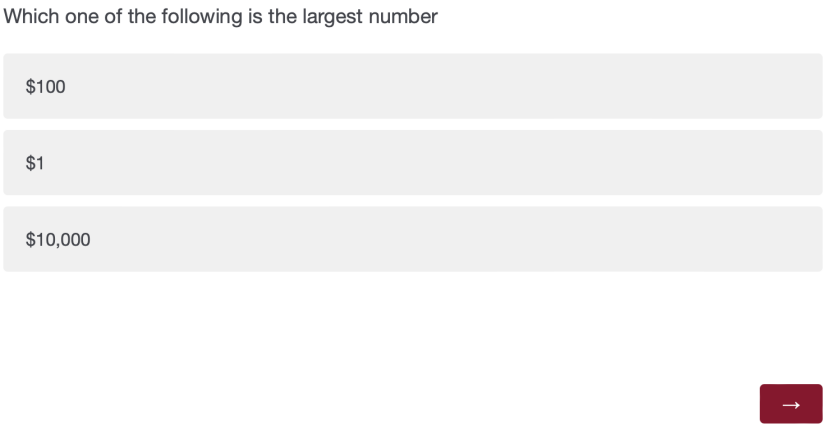


Figure LXXVI: Attention Check Question 2

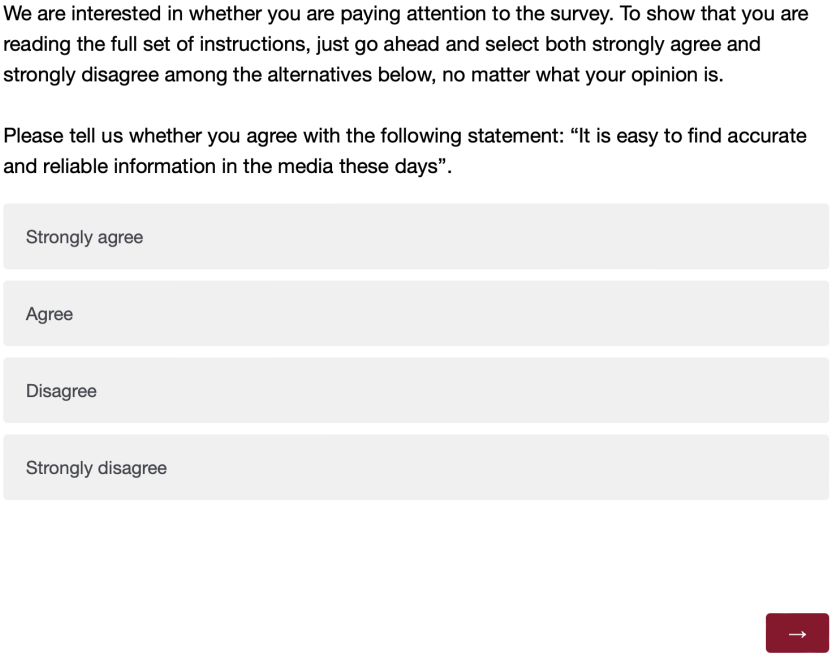


Figure LXXVII: Comprehension Question 1

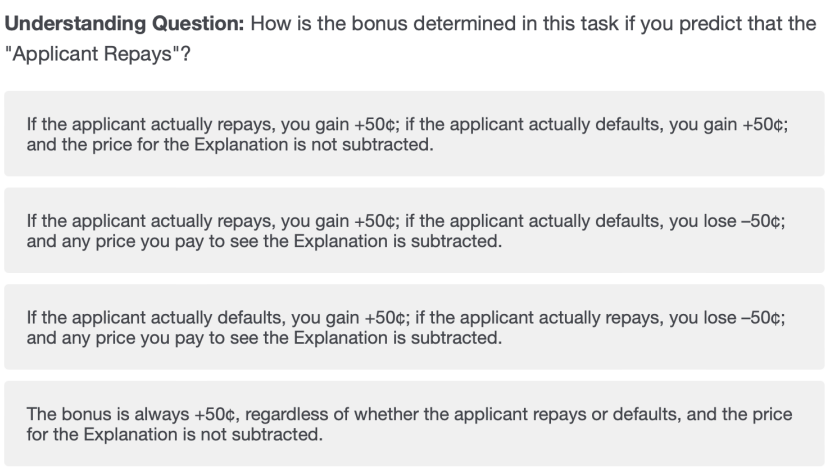


Figure LXXVIII: Comprehension Question 2

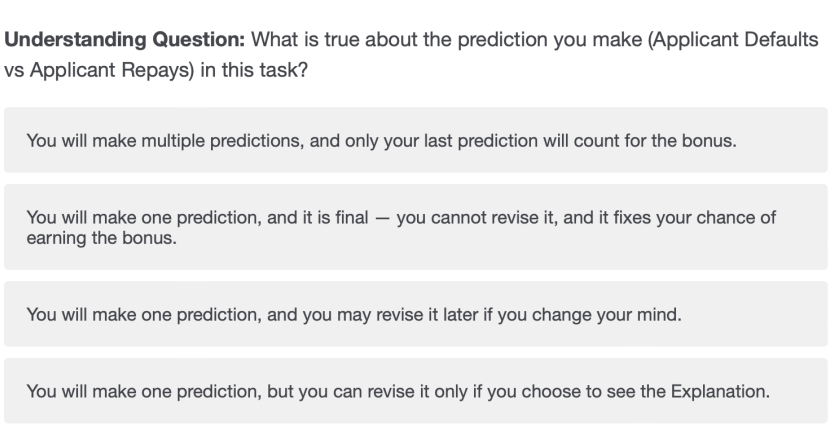


Figure LXXIX: Comprehension Question 3

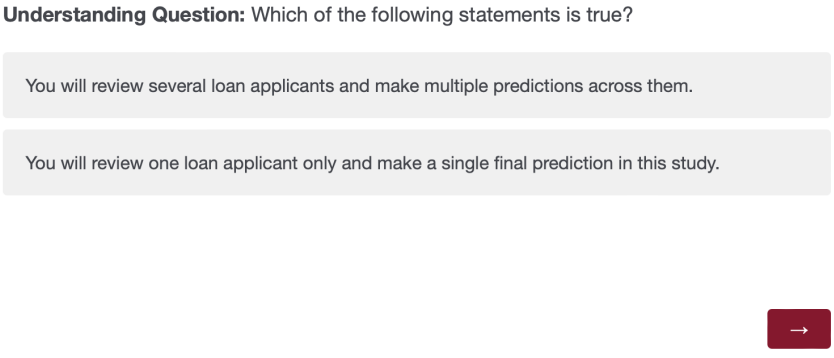


Figure LXXX: Comprehension Question 4

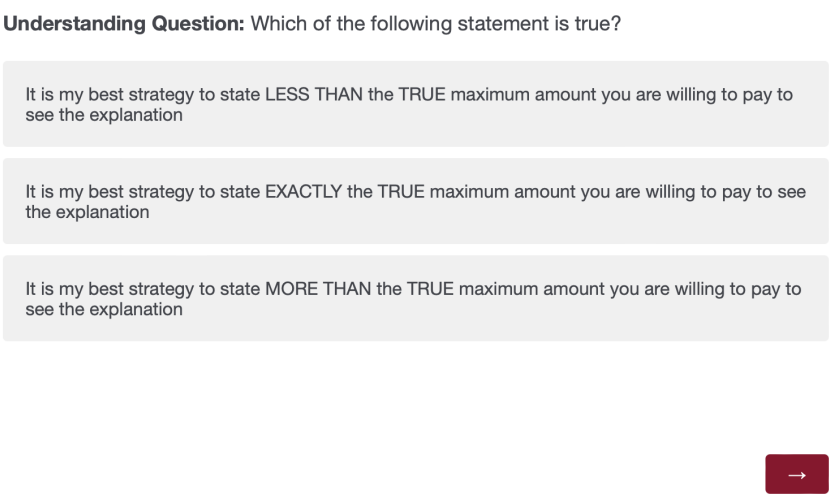


Figure LXXXI: Comprehension Question 5

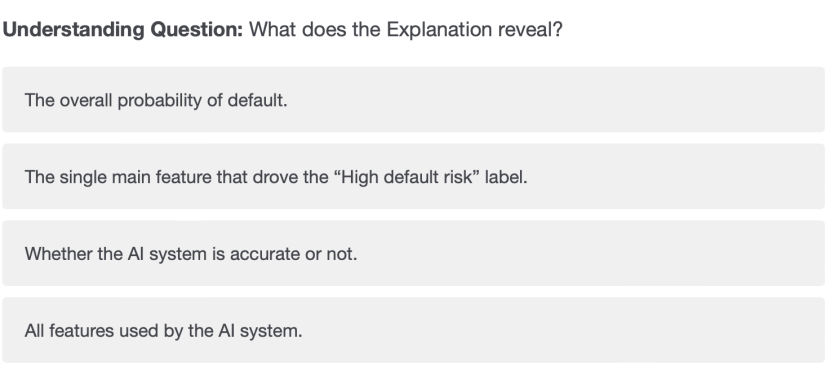


Figure LXXXII: Comprehension Question 6

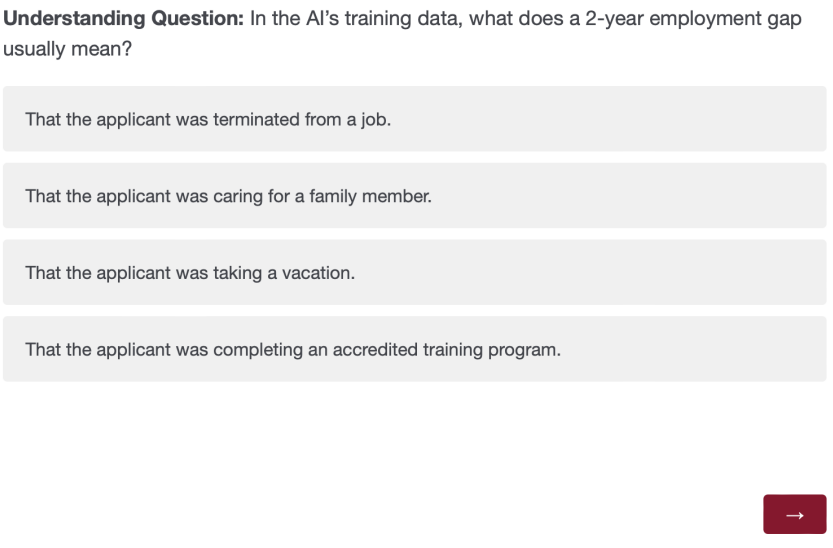


Figure LXXXIII: Comprehension Question 7

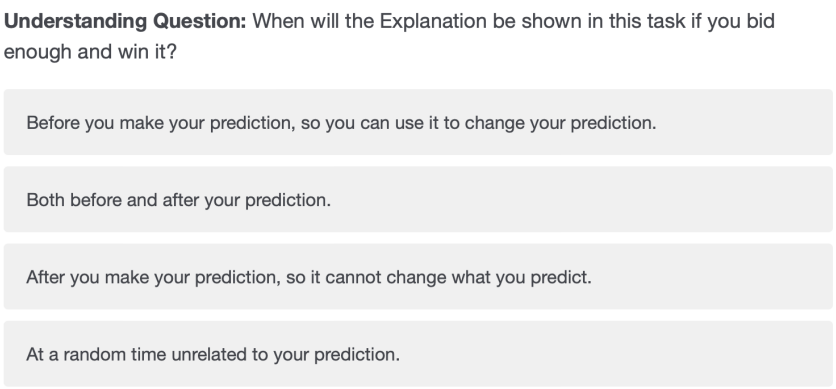


Figure LXXXIV: Comprehension Question 8

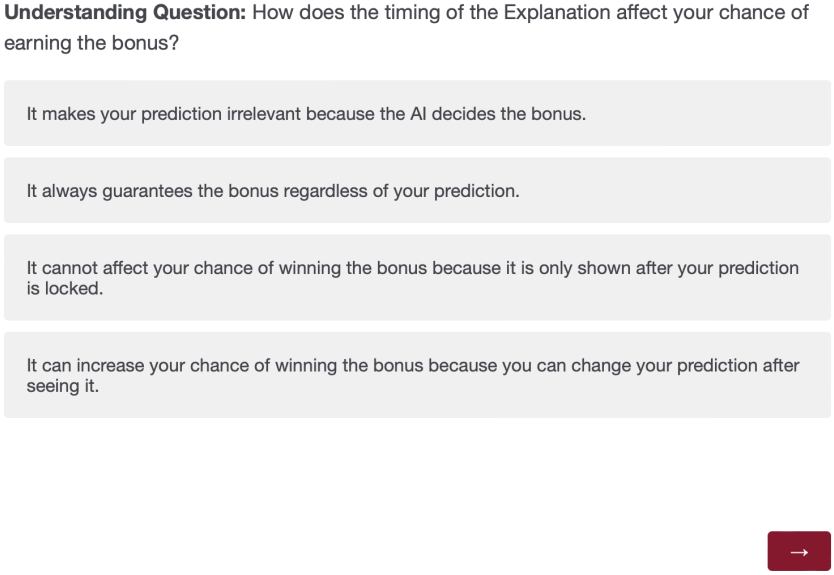


Figure LXXXV: Comprehension Question 9

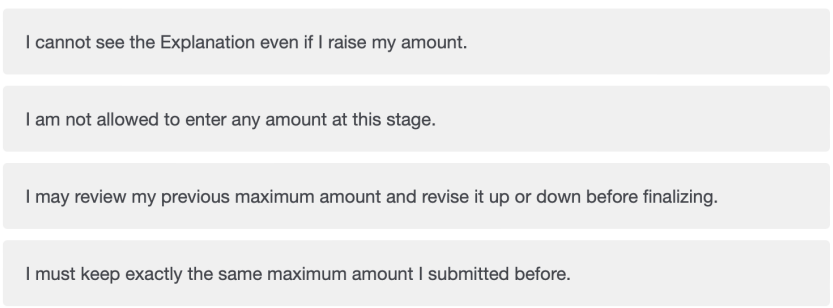


Figure LXXXVI: Comprehension Question 10

Understanding Question: How does the AI system treat a 2-year employment gap?

- It assigns lower risk when the gap is due to a certificate program.
- It ignores 2-year employment gaps entirely and never uses them.
- It distinguishes between gaps due to termination and gaps due to certificate programs.
- It always assumes that a 2-year employment gap means termination from a job.

→

Figure LXXXVII: Comprehension Question 11

Understanding Question: If the Explanation names "2-year employment gap" as the main feature, what follows in this study?

- We cannot infer anything about risk from that information.
- The applicant remains HIGH default risk (gap always means termination).
- The Explanation adds nothing; the note is irrelevant.
- The applicant should have been LOW default risk (note makes the gap benign).

Figure LXXXVIII: Comprehension Question 12

Understanding Question: If the Explanation says the main feature is not the 2-year employment gap, what follows in this study?

- The Private Note overrides any feature the Explanation names.
- The HIGH default risk label should stand (the note is about the gap and does not apply).
- We must ignore the system label entirely.
- The applicant should have been LOW default risk.

Figure LXXXIX: Comprehension Question 13

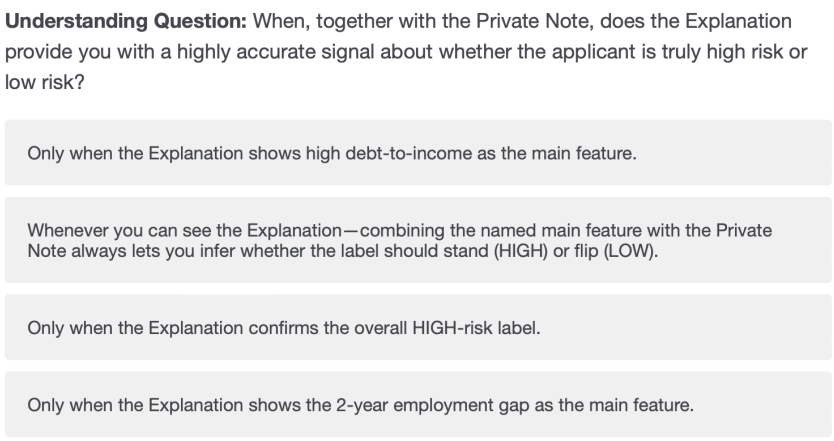


Figure XC: Comprehension Question 14 for *Before - Aided Reasoning* Condition

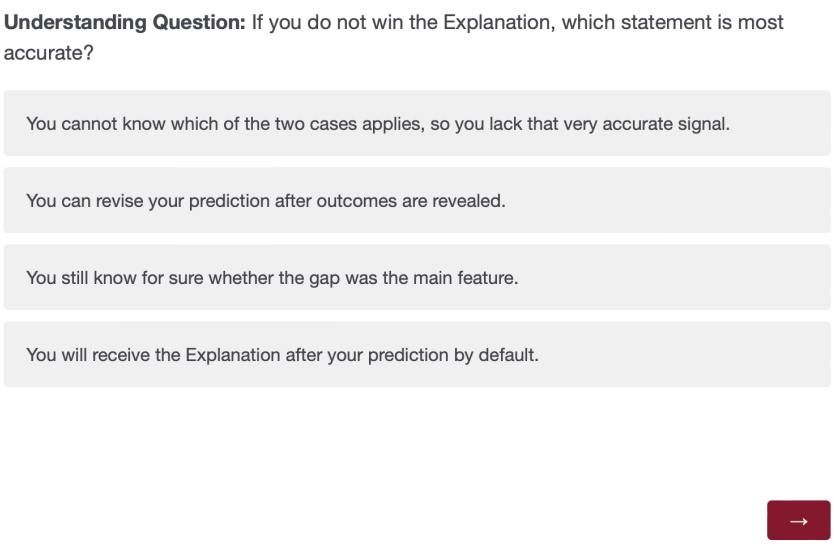
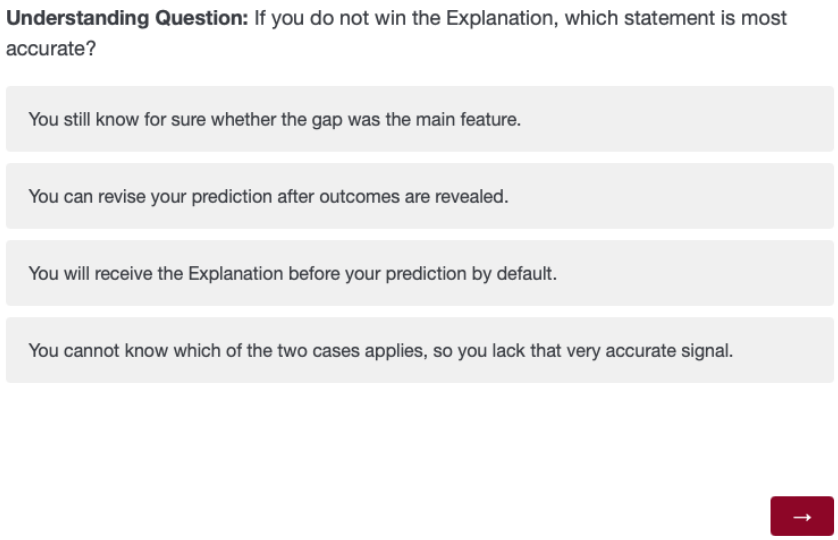


Figure XCI: Comprehension Question 14 for *After - Aided Reasoning* Condition



1.3 AI Quiz

To assess participants' baseline understanding of artificial intelligence (AI), I ask them seven AI-related questions (See Figures XCII, XCIII, XCIV, XCV, XCVI, XCVII, XCVIII). Participants in Experiment 1 are presented with all seven questions, while participants in Experiment 2 are shown only the first four questions (see Figures XCII, XCIII, XCIV, XCV).

Figure XCII: AI Question 1

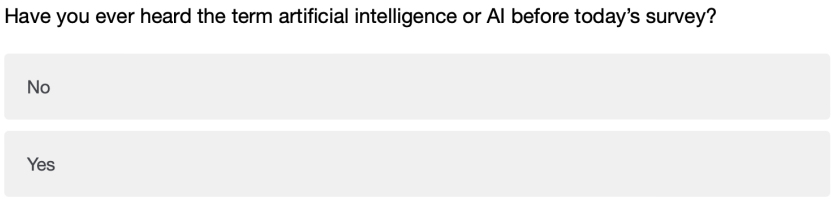


Figure XCIII: AI Question 2

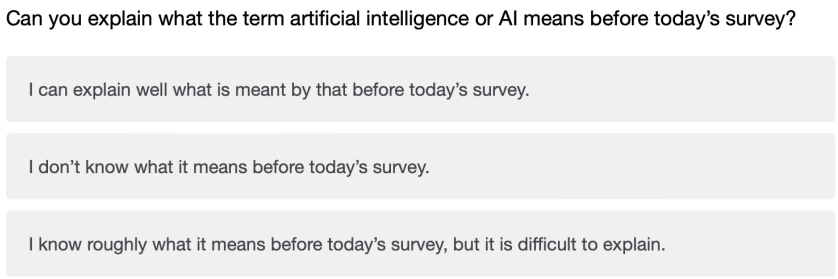


Figure XCIV: AI Question 3

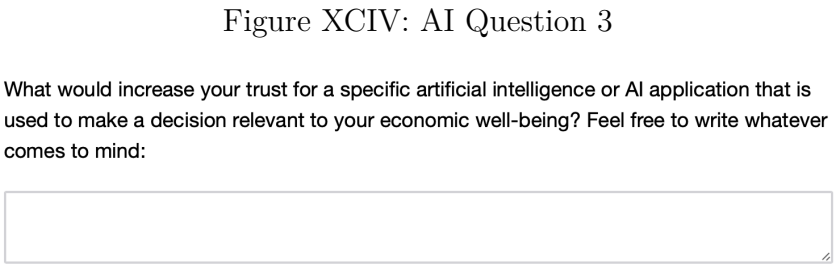


Figure XCV: AI Question 4

Figure XCVI: AI Question 5

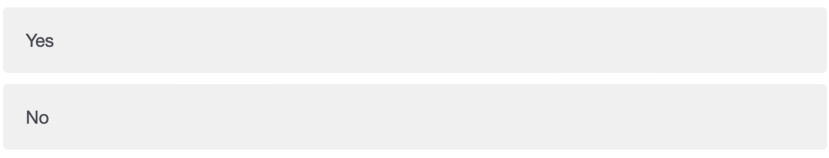


Figure XCVII: AI Question 6

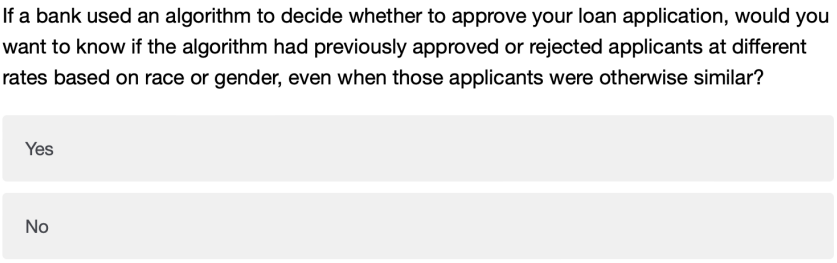
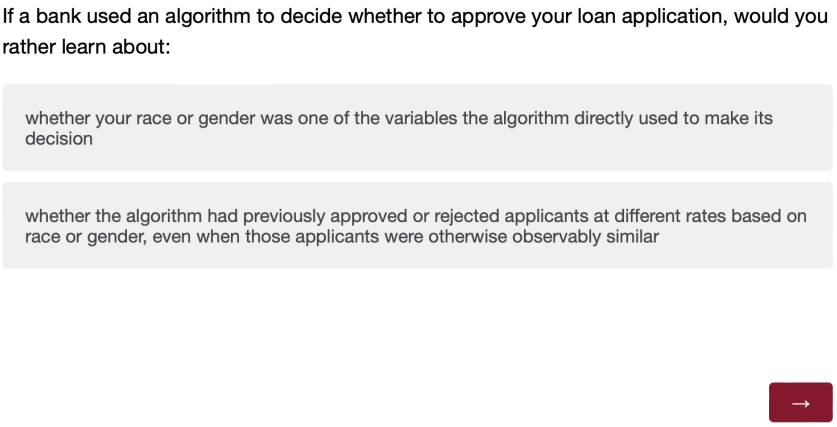


Figure XCVIII: AI Question 7



1.4 Collection of Demographic Information

In the final section of these two experiments, the participants are asked a number of questions (in the order of appearance) about their gender identity (Figure XCIX), race and ethnicity (Figure C), sexual orientation (Figure CI), disability and chronic condition (Figure CII), zip code (Figure CIII), highest education level (Figure CIV), high school type and location (Figure CV, CVI), marital status (Figure CVII), number of children (Figure CVIII), employment status (Figure CIX), occupation type (Figure CX), present religion and importance in their lives(Figure CXI, CXII), life satisfaction (Figure CXIII), health insurance (Figure CXIV), social class (Figure CXV), political affiliation (Figure CXVI), favorability toward the democratic or republican parties (Figure CXVII, CXVIII), and health status (Figure CXIX).

Figure XCIX: Survey Question about Participant's Gender Identity

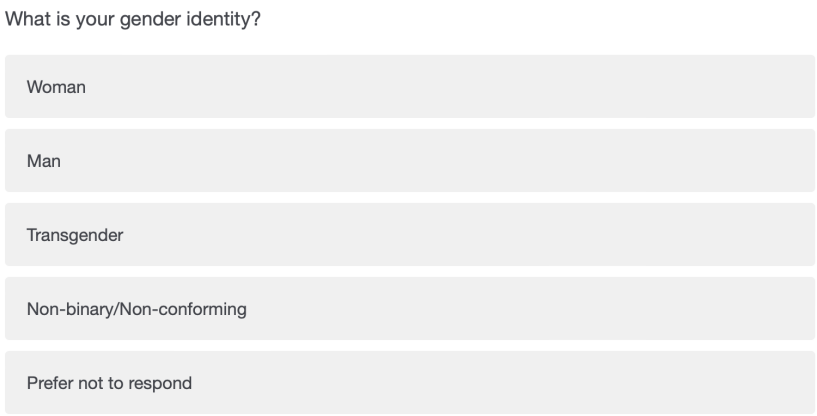


Figure C: Survey Question about Participant's Race and Ethnicity

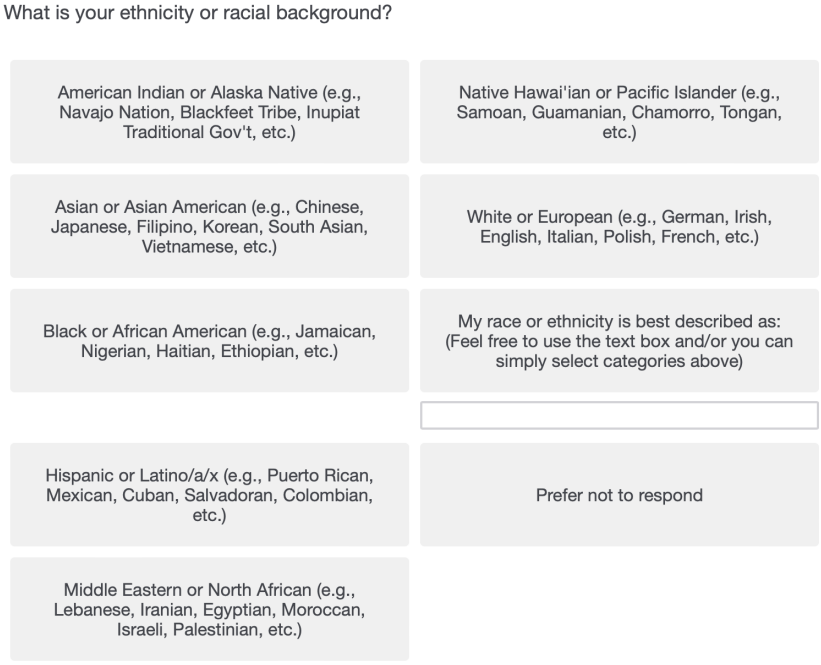


Figure CI: Survey Question about Participant's Sexual Orientation

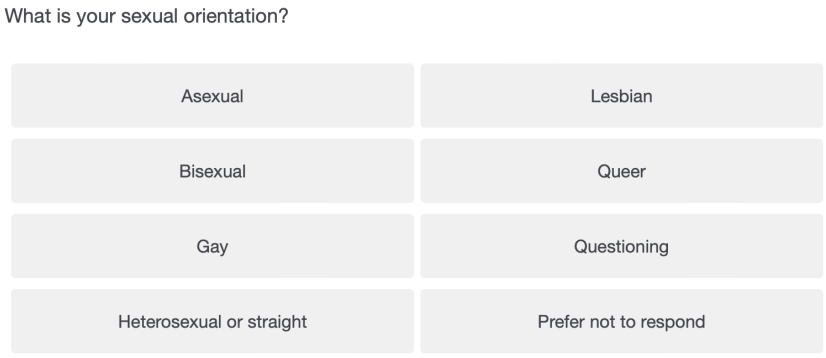


Figure CII: Survey Question about Participant's Disability and Chronic Condition

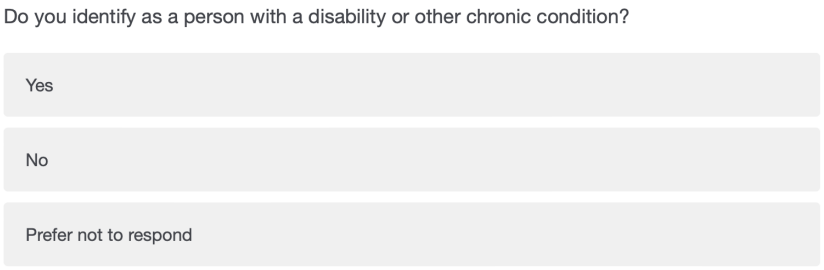


Figure CIII: Survey Question about Participant's Zip Code

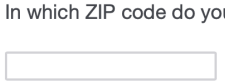


Figure CIV: Survey Question about Participant's Highest Education Level

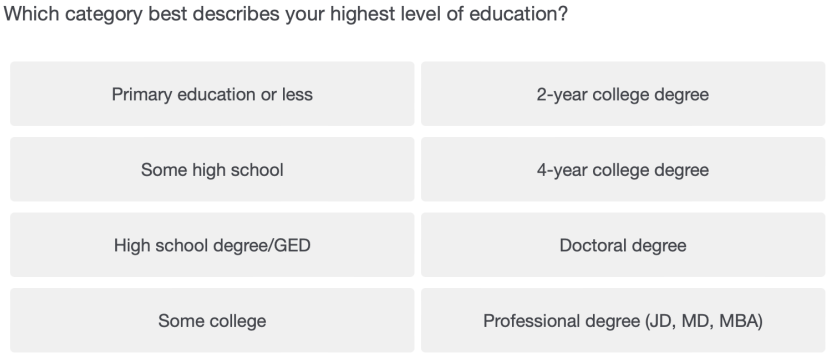


Figure CV: Survey Question about Participant's High School Type

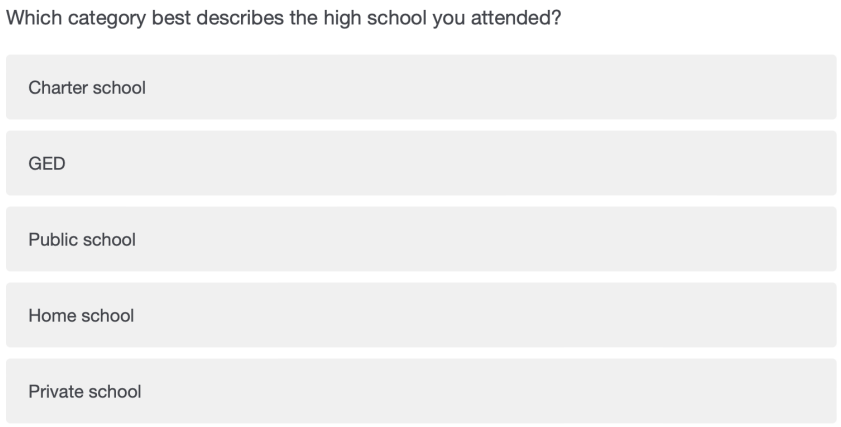


Figure CVI: Survey Question about Participant's High School Location

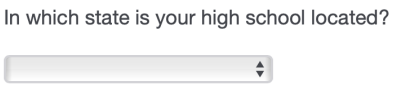


Figure CVII: Survey Question about Participant's Marital Status

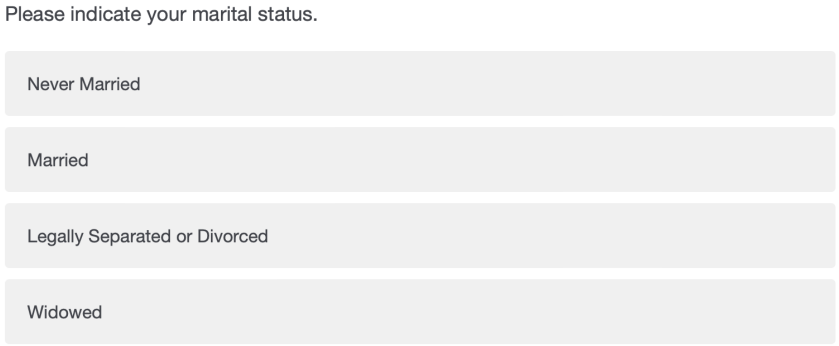


Figure CVIII: Survey Question about Participant's Number of Children

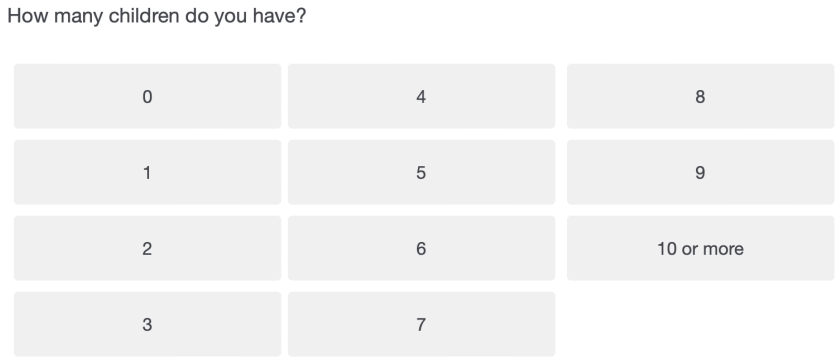


Figure CIX: Survey Question about Participant's Employment Status

Figure CX: Survey Question about Participant's Occupation Type

Figure CXI: Survey Question about Participant's Present Religion

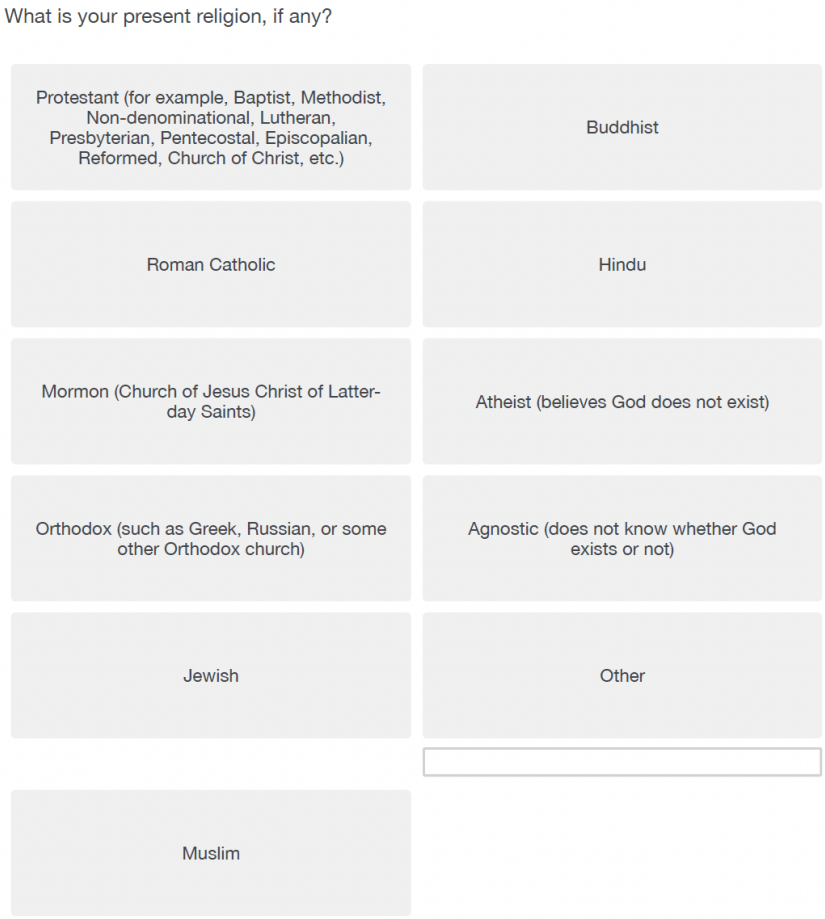


Figure CXII: Survey Question on Religion's Importance in Participants' Lives

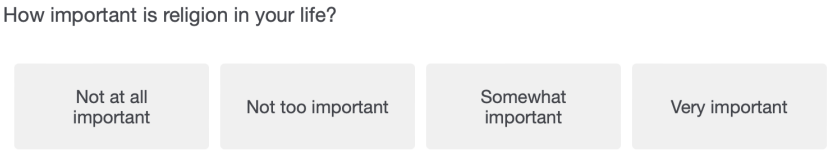


Figure CXIII: Survey Question about Participants' Life Satisfaction

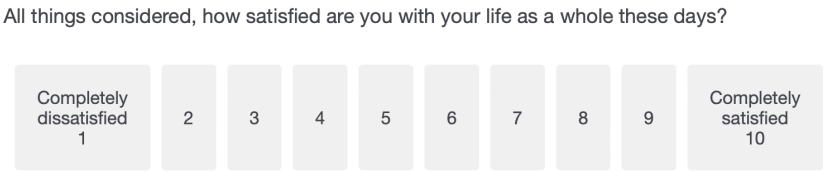


Figure CXIV: Survey Question about Participants' Health Insurance

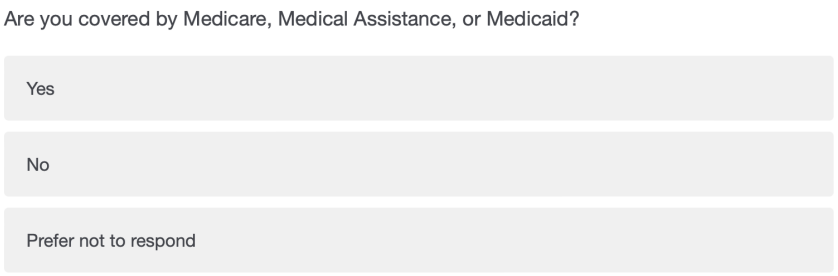


Figure CXV: Survey Question about Participants' Social Class

Figure CXVI: Survey Question about Participants' Political Affiliation

Figure CXVII: Survey Question on Participants' Favorability Toward the Democratic Party

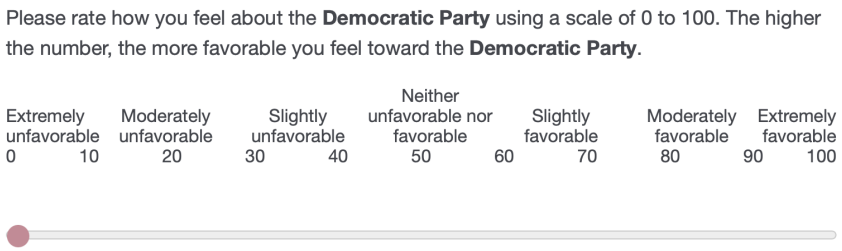


Figure CXVIII: Survey Question on Participants' Favorability Toward the Republican Party

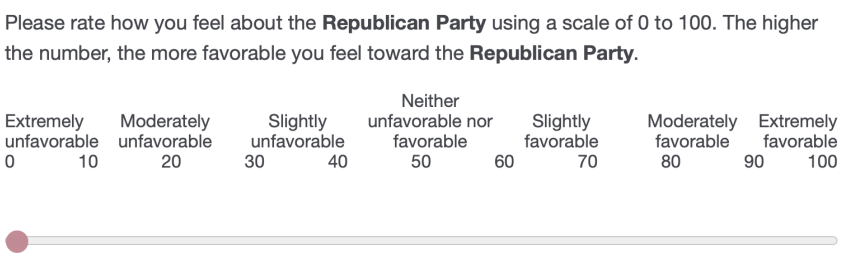


Figure CXIX: Survey Question about Participants' Health Status