

# Online Appendix for Should I State or Should I Show? Aligning AI with Human Preferences

Keaton Ellis\* Wanying Huang<sup>†</sup>

April 1, 2026

## 1 Experimental Screenshots

### 1.1 Informed Consent

---

\*Monash University ([keaton.ellis@monash.edu](mailto:keaton.ellis@monash.edu))

<sup>†</sup>Monash University ([kate.huang@monash.edu](mailto:kate.huang@monash.edu))

## Informed Consent

Please carefully review the information below and click either "I have read the information above and I consent to participate in this study" or "I do not consent to participate". This study has been approved by the Human Research Ethics Committee at Monash University (Project ID: 50731).

The full Explanatory Statement is available for download [here](#). A summary of the key information is provided below.

### Principal investigators

Dr Keaton Ellis ([keaton.ellis@monash.edu](mailto:keaton.ellis@monash.edu)) and Dr Kate Huang ([kate.huang@monash.edu](mailto:kate.huang@monash.edu)), Monash University.

### What the study is about

This research project aims to learn more about decision-making under uncertainty and how people delegate decisions to artificial intelligence (AI) systems.

### What we will ask you to do

If you choose to participate, we will ask you to make a series of choices under uncertainty and to decide whether to delegate your decisions to an AI system in similar situations. We will analyze participants' decisions in an anonymous manner (your identity will not be recorded in the dataset) to further our understanding of how people make decisions under uncertainty and when they choose to delegate decisions to an AI system.

### Duration

The study will last about 25-30 minutes.

### Compensation for participation

You will receive a participation fee of **\$5.00** for completing all parts of this study. Additionally, you will have a chance to earn additional bonus (up to **\$51.70**). How much you receive depends partly on your decisions and partly on chance. All earnings will be paid to you via Prolific.

### Privacy/Confidentiality/Data Security

We will treat all your responses in this study anonymously and use your Prolific ID for payment purposes only. Once the study is complete, we will remove this information from our data and maintain the de-identified data in a password-protected folder accessible only by the research team members. We will only report results in de-identified form. Please note that the study is hosted on an external server not affiliated with Monash University and with its own privacy and security policies. We anticipate that your participation in this study presents no greater risk than your everyday use of the Internet.

### Taking part is voluntary

Taking part in this study is voluntary, and you may withdraw from the study at any time. Withdrawing before the study is complete will mean that you will not earn a participation fee and will not be eligible for any additional compensation from this study.

### Contact Information

If you have any questions or concerns regarding your rights as a subject of this study, you may contact the Monash University Human Research Ethics Committee at [MUHREC@monash.edu](mailto:MUHREC@monash.edu)

Do you consent to participate in this study?

- I have read the information above and I consent to participate in this study.  
 I do not consent to participate.

Next

FIGURE 1: Informed Consent

## 1.2 Study Overview and Instructions

### Study Overview

Thank you for participating. Please read these instructions carefully.

#### Study Structure

This study has **three parts**. You will receive detailed instructions for each part at its start.

- **Part I** has three tasks. You will provide multiple pieces of information to help AI agents make choices between lotteries on your behalf. We will explain later in detail what each task entails.
- **Part II** has one task. You will make choices between lotteries.
- **Part III** asks you to answer additional survey questions and complete a short reasoning test.

#### Payment

You will receive:

- **Base payment:** \$5.00 for completing the study
- **Lottery bonus:** We will randomly select a choice from all choices made during Parts I and II. The selected choice might have been made by you, or might have been made by an AI agent on your behalf. The selected choice (i.e. a lottery) in the question will then be played out for real money. The highest possible lottery outcome is \$47.00.
- **Survey bonus:** Some survey questions in Part III involve bonus payments. We will explain later in detail how the bonuses will be determined.

**Please complete each task carefully**, as choices from each task could be selected for your payment.

While answer the questions, we ask you to **not browse the internet** and **not consult with others**. We are genuinely interested in your opinions and preferences.

Next

FIGURE 2: Study Overview

## Part I Instructions

You are now reading the instructions for **Part I**. In Part I, you will work with lotteries and help AI agents make lottery choices on your behalf.

### Understanding Lotteries

A **lottery** is a gamble with uncertain outcomes. Each lottery shows you the possible amounts you could win and the probability (chance) of each outcome.

Throughout this study, you will choose between pairs of lotteries presented like this:

Lottery A	
Probability	Outcome
50%	\$4.00
50%	\$0.00

Lottery B	
Probability	Outcome
100%	\$1.50

**Lottery A** means there is a 50% chance you win \$4.00 and a 50% chance you win \$0.00.

**Lottery B** means you win \$1.50 for certain.

When you choose between two lotteries, you are selecting which gamble you prefer to play.

Next

FIGURE 3: Part I Instructions: Understanding Lotteries

## Part I Instructions

### Additional Information About Lotteries

When picking, you will see five additional pieces of information:

**Average Payment:** (also known as "expected value") This is the average payment the lottery would pay out if it were played many, many times. Lotteries with higher average payments pay more on average, but there is still randomness to how much they pay each time. To calculate the average payment, we multiply each outcome by the probability of that outcome occurring and add this up for all the outcomes.

**Payment Variability:** (also known as "variance") This is a measure of how much payments can vary. Lotteries with higher payment variability pay amounts that are more spread out, which often means the difference between the larger and lower payment is larger. To calculate the payment variability, we subtract each outcome from the average payment, square it, and then add this up for all the outcomes, weighted by the respective probabilities.

**Minimum Payment:** This is the minimum possible amount the lottery could pay.

**Maximum Payment:** This is the maximum possible amount the lottery could pay.

**Chance of Max Payment:** This is the probability that the lottery pays the Maximum Payment.

We will show you two lotteries, and you can access this additional information about them by clicking on the buttons at the bottom, as we show below. When you click on a button, it will display that information for both lotteries:

Lottery A		Lottery B	
Probability	Outcome	Probability	Outcome
50%	\$4.00	100%	\$1.50
50%	\$0.00		

Additional information:

Please click on the buttons above to get a sense of how this works for the example lotteries above.

When making your choices between lotteries, you will be able to click to learn as many pieces of additional information you want about the lotteries. If you wish, you can use this information to help you make your choices. **Please note that you do not have to use these buttons at all; they are just for your convenience.**

When you are ready, continue.

FIGURE 4: Part I Instructions: Additional Information About Lotteries

## Comprehension Quiz

Attempt 1 of 3

Please answer all questions correctly to continue. You have 3 attempt(s) remaining.

### Question 1:

How will your bonus payment be determined?

- The outcomes of all lottery choices will be added together.
- One lottery choice will be randomly selected and played out.
- The lottery with the highest expected value will be played.

### Question 2:

If a lottery has a 50% chance of \$4.00 and 50% chance of \$0.00, what are the possible outcomes?

- You could win either \$4.00 or \$0.00
- You will always win \$2.00 (the average)
- You will always win \$4.00

### Question 3:

When writing instructions for the AI, what should you aim for?

- Write instructions that make the AI choose randomly.
- Write instructions that lead the AI to make choices you prefer.
- Write instructions that make the AI respond as quickly as possible.

### Question 4:

What is the maximum total bonus you can earn?

- \$1.00
- \$10.00
- \$51.70

Next

FIGURE 5: Comprehension Quiz

## 1.3 Part I, Task 1: Initial Lottery Choices

### Part I, Task 1: Initial Lottery Choices

In this task, you will make **13 choices** between pairs of lotteries. In each choice, you will see two lotteries with different probabilities and potential payoffs. Please select the lottery you prefer to play.

Your choices will be shown to an AI agent (Claude, developed by Anthropic). Based on your **13 choices**, the AI agent will then make choices on your behalf between **new pairs of lotteries that are similar in structure but not identical** to those in this task. Because one of the AI agent's choices may be used to determine your bonus payment, it is important that you choose the lotteries you truly prefer so that the AI agent can learn your preferences accurately.

Next

FIGURE 6: Part I, Task 1 Introduction

### Question 1 of 13

Click on the lottery you prefer:

Lottery A

Probability	Outcome
20%	\$4.00
80%	\$0.00

Lottery B

Probability	Outcome
25%	\$3.00
75%	\$0.00

Click to compare statistics:

Average PaymentPayment VariabilityMinimum PaymentMaximum PaymentChance of Max

Next

FIGURE 7: Lottery Choice Screen (Part I)

**Can you spot the animal camouflaged in the image below?**

Please click on the photo where you think the animal is located

This task does not affect your payment and is just for fun.



This photo was originally published by the Reader's Digest on August 22, 2019 (<https://www.rd.com/list/camouflaged-animal-photos/>)

Next

FIGURE 8: Attention Break

## 1.4 Part I, Task 2: Writing Instructions for an AI

### Part I, Task 2: Writing Instructions for an AI

In this task, you will **write instructions** to an AI agent (Claude, developed by Anthropic). Based on your instructions, the AI agent will then make choices on your behalf between **new pairs of lotteries that are similar in structure but not identical** to those shown in Task 1. Because one of the AI agent's choices may be used to determine your bonus payment, it is important that you write the instructions carefully so that the AI agent can learn your preferences accurately.

**Note:** Claude will receive your instructions, as well as some details about the task. [Click here to see the full prompt we will send to Claude.](#)

Next

FIGURE 9: Part I, Task 2: Writing Instructions for an AI

### Write Your Instructions for Claude

Please spend at least 1 minute writing your instructions. Time remaining: 59 seconds

**Note:** Good instructions will provide clear, flexible guidance that can be used across different situations. Bad instructions are too specific to a particular scenario and don't adapt to changing conditions.

You might find it helpful to reference your choices in Task 1 when writing your instructions. [Click here to see your previous choices.](#)

#### Write your instructions for Claude:

Write clear instructions that describe your preferences for choosing between lotteries. Please write your prompt carefully, as Claude will use these instructions to make choices on your behalf in **new questions**.

While writing your instructions, we ask you to **not browse the internet** and **not consult with others**. We are genuinely interested in your opinions and preferences.

Next

FIGURE 10: Write Your Instructions for Claude

## 1.5 Part I, Task 3: Delegation Choice

### Part I, Task 3: Delegation Choice

In this task, Claude will again use the information you provided to make choices on your behalf between **13 new pairs of lotteries (Questions 14–26)**. This time, you will choose what information to send to Claude.

#### Your Delegation Choice

You can choose **exactly one** (not both) of the following two pieces of information to send to Claude. Once you choose one option, the other option will not be used by Claude.

- **Written instructions (prompt):** If you choose this option, Claude will use the instructions you wrote in Task 2 to infer your preferences and make choices on your behalf for Questions 14–26.
- or
- **Raw choice data:** If you choose this option, Claude will observe your lottery choices from Questions 1–13 in Task 1 and use them to infer your preferences and make choices on your behalf for Questions 14–26.

**Next:** You will review the information available and make your delegation choice.

Next

FIGURE 11: Part I, Task 3: Delegation Choice

### Part I, Task 3: Delegation Choice

#### What Would You Like to Send to the AI?

Review your information above, then click on your choice below:

##### Your Written Instructions:

"[SUBJECT INSTRUCTIONS HERE]"

##### Your Task 1 Choice Data:

Show Details

13 lottery choices recorded

Click on your choice:

##### Written Instructions Only

Send only your written instructions to the AI

##### Choice Data Only

Send only your lottery choice data to the AI

Next

FIGURE 12: Delegation Choice Screen

## 1.6 Part II: Your Lottery Choices

### Part II: Your Lottery Choices

This is the second part of the study, and it contains one task.

In this task, you will make your own choices between **13 new pairs of lotteries**, which are **different from those you saw in Part I**. For each choice, select the lottery you prefer.

Please think about each choice carefully, as one of your choices may be selected to determine your bonus payment.

Next

FIGURE 13: Part II: Your Lottery Choices

### Question 14 of 26

Click on the lottery you prefer:

#### Lottery A

Probability	Outcome
95%	\$3.00
5%	\$0.00

#### Lottery B

Probability	Outcome
91%	\$4.00
9%	\$0.00

Click to compare statistics:

Average Payment Payment Variability Minimum Payment Maximum Payment Chance of Max

Next

FIGURE 14: Lottery Choice Screen (Part II)

**Can you spot the animal camouflaged in the image below?**

Please click on the photo where you think the animal is located

This task does not affect your payment and is just for fun.



This photo was originally published by the Reader's Digest on August 22, 2019 (<https://www.rd.com/list/camouflaged-animal-photos/>)

Next

FIGURE 15: Attention Break

## 1.7 Part III: Exit Survey

### Part III Instructions

This is the third part of the study. In this part, you will answer some short questions that are unrelated and different from all previous questions.

This part is short and should take you about 5-10 minutes to complete.

There are **no right or wrong answers**. We are interested in studying **your preferences**.

In addition to the potential bonus from Parts I and II, you may earn an **extra bonus** in this part. We will explain later how this is determined.

Next

FIGURE 16: Part III Instructions

## Part III: Your Decision-Making Reasoning

### Your Reasoning

Earlier in this experiment, you chose to send **written instructions (prompt)** to the AI to help it make decisions on your behalf.

Why did you make this choice? Please explain your reasoning:

### Predict AI Performance

**Bonus opportunity:** You can earn \$0.25 for each correct prediction below (up to \$0.50 total). This bonus is in addition to your other earnings.

**AI with your written instructions (prompt):**

If the AI were to answer the same questions (Questions 14-26) on your behalf based on your written instructions, how many of them do you think the AI would have answered the same way as you did?

**AI with your lottery choice data:**

If the AI were to answer the same questions (Questions 14-26) on your behalf based only on your lottery choices from Questions 1-13, how many of them do you think the AI would have answered the same way as you did?

FIGURE 17: Part III: Your Decision-Making Reasoning

## Part III: Pattern Matching Test

This part of the study has **6 graphical puzzles**.

Each puzzle you solve correctly pays **\$0.20**.


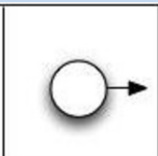

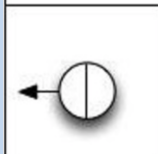
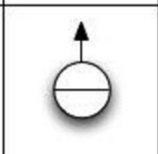
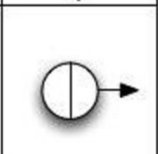
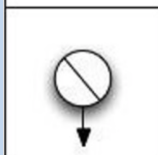
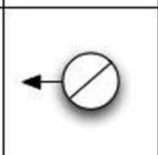

If you answer all 6 puzzles correctly, you can earn up to **\$1.20**.

For each puzzle, you will see a pattern with a missing piece. Select the answer that best completes the pattern.

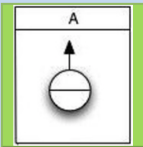
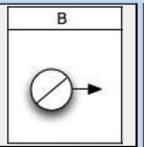
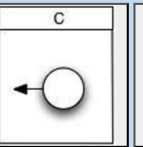
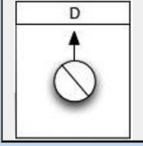
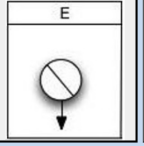
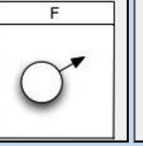
Next

FIGURE 18: Pattern Matching Test Introduction

**Graphical Puzzle 1**

Which one comes next?

<b>A</b> 	<b>B</b> 	<b>C</b> 	<b>I don't know</b>
<b>D</b> 	<b>E</b> 	<b>F</b> 	<b>None of these</b>

Next

FIGURE 19: Pattern Matching Test: Example Puzzle

### Part III: Pattern Matching Test

How many of the 6 previous puzzles do you think you answered correctly?

If that number coincides with the actual number of puzzles you solved correctly, you will receive 25 cents. Otherwise, you will receive 0 cents for this question.

Next

FIGURE 20: Pattern Matching Test: Self-Assessment

### Part III: Pattern Matching Test

We ran a similar experiment on Prolific and have data from 100 random Prolific participants. How many of these 100 correctly solved more puzzles than you?

Please enter a number between 0 and 100:

Next

FIGURE 21: Pattern Matching Test: Relative Performance

## Part III: Investment Decision (1 of 2)

You are endowed with **100 tokens** (worth **\$0.50**) and asked to choose the portion of this amount (between 0 and 100 tokens, inclusive) that you wish to invest in a risky option. Tokens not invested are yours to keep.

**If the investment is successful:** You receive **2.5x** the amount you invested.  
**If the investment is unsuccessful:** You lose the amount invested.

To determine if the investment is successful or not, we will flip a fair coin:

**Heads (50%): Success**      **Tails (50%): Failure**

The coin flip will be performed after you complete the study, and your bonus will be determined accordingly.

**How many tokens do you wish to invest?**

0 (keep all)      100 (invest all)

I wish to invest this many tokens:

**Your potential outcomes:**

**If successful:** You will have **100 tokens** (\$0.50)  
**If unsuccessful:** You will have **100 tokens** (\$0.50)

[Next](#)

FIGURE 22: Investment Decision 1 of 2

## Part III: Investment Decision (2 of 2)

You are endowed with **100 tokens** (worth **\$0.50**) and asked to choose the portion of this amount (between 0 and 100 tokens, inclusive) that you wish to invest in a risky option. Tokens not invested are yours to keep.

**If the investment is successful:** You receive **2.5x** the amount you invested.  
**If the investment is unsuccessful:** You lose the amount invested.

To determine if the investment is successful or not, we will roll a **four-faced die** with faces marked A, B, C, D:

**First Roll:**

- A** = Investment is **successful**
- D** = Investment is **unsuccessful**
- B** or **C** = Roll the die again

**Second Roll (if B or C on first roll):**

- A** or **B** = Investment is **successful**
- C** or **D** = Investment is **unsuccessful**

The die rolls will be performed after you complete the study, and your bonus will be determined accordingly.

**How many tokens do you wish to invest?**

0 (keep all) 100 (invest all)

I wish to invest this many tokens:

**Your potential outcomes:**

- If successful:** You will have **100 tokens** (\$0.50)
- If unsuccessful:** You will have **100 tokens** (\$0.50)

Next

FIGURE 23: Investment Decision 2 of 2

## Part III: Background Survey

Please answer the following questions about yourself. Your responses will help us understand how different backgrounds relate to decision-making preferences.

### AI Experience

How comfortable are you with using AI tools (e.g., ChatGPT, Claude, Copilot)?

Not at all comfortable Neutral Very comfortable

1    2    3    4    5    6    7

### AI Usage

How often do you use AI tools (e.g., ChatGPT, Claude, Copilot) in your daily life or work?

Never  
 Rarely (a few times a year)  
 Monthly  
 Weekly  
 Daily

### Writing

How comfortable are you with writing instructions or explanations for others?

Not at all comfortable Neutral Very comfortable

1    2    3    4    5    6    7

### Professional Background

Do you have experience managing or supervising other people?

No management experience  
 Some experience (1-2 years)  
 Moderate experience (3-5 years)  
 Extensive experience (6+ years)

### Education

What is the highest level of education you have completed?

High school or equivalent  
 Some college, no degree  
 Associate's degree  
 Bachelor's degree  
 Master's degree  
 Doctorate or professional degree

### Personal Traits

How would you rate your level of impatience in general?

Very patient Neutral Very impatient

1    2    3    4    5    6    7

[Next](#)

FIGURE 24: Background Survey



## Your Results

### Experiment Complete

Thank you for completing this study. The payments described below will be sent through Prolific.

#### Your Payment

Base Payment:	\$5.00
Lottery Bonus:	\$1.00
Prediction Bonus:	\$0.00
Graphical Puzzle Bonus:	\$0.20
Investment Bonus:	\$1.00
<b>Total Payment:</b>	<b>\$7.20</b>

#### How Your Lottery Bonus Was Determined

**Selected from:** AI's Choices (Questions 14-26, using your choice data)

**Question:** 25

**Choice:** Lottery A

**Lottery played:** 25% chance of \$1.00, 25% chance of \$3.50, 25% chance of \$6.00, 25% chance of \$9.25

**Outcome:** You won \$1.00 from the lottery.

[View All Lottery Choices](#)

#### Your Prediction Results

You predicted how many of the 13 questions (Q14-26) the AI would answer the same as you:

AI Method	Your Guess	Actual Matches	Bonus
AI with prompt	4	7	\$0.00
AI with choice data	2	6	\$0.00

#### Your Graphical Puzzle Results

You solved 1 out of 6 puzzles correctly.

Component	Bonus
Puzzles correct (1 x \$0.20)	\$0.20
Guessing own score	\$0.00
Guessing relative performance	\$0.00

#### Your Investment Decision Results

Both of your investment decisions count towards your payment.

##### Decision 1: Simple Coin Flip

You invested 0 tokens out of 100.

Tokens kept	100
Investment outcome (Success)	+0.0
<b>Total: 100.0 tokens</b>	<b>\$0.50</b>

##### Decision 2: Two-Stage Die Roll

You invested 0 tokens out of 100.

Tokens kept	100
Investment outcome (Unsuccessful)	0
<b>Total: 100 tokens</b>	<b>\$0.50</b>

**Total Investment Bonus: \$1.00 (\$0.50 + \$0.50)**

Please click [here](#) to complete the study and return to Prolific. Once you do, you may close this window.

FIGURE 26: Payment Summary