



## ABSTRACT

The nature of trust relationships between users and the many computerized agents people interact with daily is poorly understood. This experiment explored how such trust relationships are built and broken by asking study subjects to choose between two suggestions made by two agents in a game setting to explore if and how such trust relationships form in a real world setting.

## INTRODUCTION, MOTIVATION, AND BACKGROUND

Today’s computerized agents are responsible for everything from banking and airplane navigation, to choosing which songs to suggest to you. They’re able to anticipate how people behave with astounding accuracy in many situations. However, with the growing ubiquity of ‘smart’ or otherwise anticipatory software in our lives, how we interact with them is in need of significantly more research.

At the surface level, trust is a concept often reserved for the living, however, at a smaller scale, *functional trust* relationships are plentiful, existing between people and the many systems we interact with and allowing us to achieve greater utility from the technologies we use daily.

**Functional Trust** is trust that something (or someone) can perform a task.<sup>1</sup>

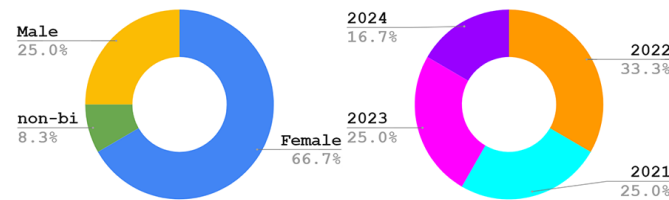


FIGURE 1: (Left) Gender distribution of subjects. (Right) Class year distribution of subjects.

## QUESTION

How much low-quality advice will a user tolerate from an agent before they distrust it?

## INTERFACE

A modified match-3 style game written in Java using the swing GUI (see *figure 2*) was used for the experiment. Similar to the popular Candy Crush Saga or Bejeweled, the interface presented 5 set scenarios to the user which each allowed the user to play a 4-move match-3 game with one major modification: users were forced to choose between to preselected ‘swaps’ presented by opposing computerized agents.

Additionally the two agents had the ability to notify the user when they successfully scored points in the form of a complement and when they missed out on a better move as a complaint from the non-chosen agent (see *figure 3*).

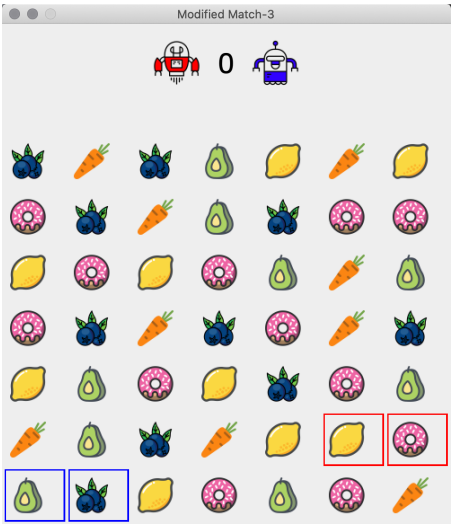




FIGURE 2: The modified interface displaying two agent’s suggestions (outlined in blue and red).

## EXPERIMENT METHODS



Due to COVID-19 restrictions, experiments were held via zoom meeting, taking advantage of the zoom remote control feature to allow subjects to play. Venmo and PayPal were used to facilitate payments for the same reason. Demographics of the 12 subjects can be found in *figure 1*.

To motivate subjects to make thoughtful choices throughout the game, subjects were paid \$6 for their time, with the opportunity to earn up to \$10 by earning above a preset point threshold during the 4 non-tutorial rounds.



## RESULTS

Scenario 2		
Move 1	6 (cascade)	3 (single)
Move 2-4	3 (single)	6 (cascade)



**-Blue choosers swap to red after first move**  
5/7 Participants chose to swap, despite earning a high value move by choosing blue in move 1.  
**-Red choosers swap to blue after first move**  
All 5 participants swapped (agent notified subject).

Scenario 3		
Moves 1 - 4	3 (single)	6 (cascade)

**-Only 1/5 blue choosers chose red move 2**  
Despite complaints each round, 2/5 waited until move 3, the last 2/5 swapped in the last move.  
**-Red choosers stayed with red for 2 moves**  
4/7 Chose red the entire game while 3 tried blue during the final move.

Scenario 4		
Move 1	3 (single)	3 (single)
Move 2	7 (cascade)	3 (single)
Moves 3-4	3 (single)	4 (single)

**-Move 2 red choosers didn’t reliably swap**  
Despite complaints, subjects had no preference.  
**-All move 2 blue choosers stayed with blue**  
The 4 subjects who chose blue remained with blue for the rest of the round, indicating a possible trust relationship between the subject and the blue agent.

Scenario 5		
Move 1	9 (cascade)	9 (cascade)
Move 2-4	3 (single)	3 (single)

**-Agents deceptively complain moves 2-4**  
**-Half of subjects remained with initial choice**  
Those 6 stuck with their first choice the whole game.  
**-Only 42% of complaints lead to a swap**  
During this round, complaints were not effective.

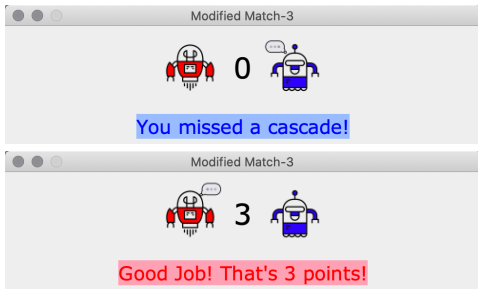


FIGURE 3: (Top) The blue agent complains that it’s move wasn’t chosen. (Bottom) The red agent congratulates the user for points scored.

### Impact of withheld payment:

When subjects missed out on part of their compensation by under-performing during one of the main rounds, they were significantly more likely to take more time to complete the following scenarios, and switch agents on a first notification.

### Summary:

Study subjects were observed making choices that violated many of the presumptions made about the effects of complaints and complements from an agent. Conversely, many of the choices made indicate minimal trust relationship formation at a reportable level.

## FUTURE WORK

### Larger Group Size:

The main limitation of the findings is lower than anticipated number of study subjects. Any future work concerning the trust relationships studied here would benefit from a much larger sample size.

### Larger Game Scenarios:

For consistency across experiment sessions each game scenario was entirely pre-decided in the design phase. As such, each consisted of a binary tree of game states which were hand arranged. Future iterations on this experiment design would benefit from deeper binary trees containing more moves. With longer games there is undoubtedly more time for trust relationships to develop, but such scenarios require a significantly larger number of scenarios, and thus: more time.

## REFERENCES

① Jøsang, A. et al. “Trust network analysis with subjective logic.” ACSC (2006).

# Trust Response to Anticipatory Software Agents

Henry Jones

June 9, 2021

## Abstract

**Purpose** - The nature of trust relationships between users and the many computerized agents people interact with daily is poorly understood. The results of this experiment will be used to gather insight into the threshold at which a user will stop following an agents advice based on the quality of its anticipation relative to another agent's anticipation.

**Methodology** - This experiment explored how trust relationships are built and broken by forcing study subjects to choose between two suggestions made by two agents of varying reliability in a game setting, simulating relative levels of anticipation of future game states. The user was tasked with deciding which suggestion to follow. During the experiment the participant's choices were logged along with the value of their chosen agent's suggested move in relation to the opposing agent's.

**Hypothesis** - If a computerized agent does not effectively anticipate the needs of its user than the user will lose trust in that agent.

**Results** - Evidence of functional trust relationships and their effects on gameplay were observed after implementing the experiment with a group of 12 study subject, results are conjectural due to the small sample size, but point towards the existence of certain conditions that may induce trust relationships between participants and computerized agents in a replicable manor.

# Contents

<b>1</b>	<b>Background</b>	<b>2</b>
<b>2</b>	<b>Related work</b>	<b>2</b>
<b>3</b>	<b>Methods and Design</b>	<b>3</b>
3.1	Summary . . . . .	3
3.2	Goals . . . . .	3
3.3	Cost . . . . .	3
3.3.1	Compensation . . . . .	3
3.4	Safety . . . . .	3
3.5	Interface Design . . . . .	4
3.5.1	Game Requirements . . . . .	4
3.5.2	Anticipatory Agent . . . . .	4
3.5.3	Scenarios . . . . .	5
3.5.4	Analytical Tools . . . . .	5
<b>4</b>	<b>Implementation</b>	<b>6</b>
4.1	Summary . . . . .	6
4.2	Game interface . . . . .	6
4.3	Scenario Building . . . . .	6
4.3.1	Implemented scenarios . . . . .	7
4.4	Study subjects . . . . .	8
4.5	Online testing venue . . . . .	8
4.6	Protocol . . . . .	8
<b>5</b>	<b>Results</b>	<b>9</b>
5.1	Summary . . . . .	9
5.2	Observations by scenario . . . . .	9
5.2.1	Scenario 2 . . . . .	9
5.2.2	Scenario 3 . . . . .	9
5.2.3	Scenario 4 . . . . .	10
5.2.4	Scenario 5 . . . . .	10
5.3	Effect of withholding payments . . . . .	11
5.4	User response to past knowledge . . . . .	11
<b>6</b>	<b>Validity threats</b>	<b>11</b>
6.1	Inadequate notifications . . . . .	11
6.2	Insufficient graphical and audio elements . . . . .	12
<b>7</b>	<b>Ethical concerns</b>	<b>12</b>
7.1	Intentional choice manipulation in the real world . . . . .	12
<b>8</b>	<b>Future work</b>	<b>12</b>
8.1	Larger sample size . . . . .	12
8.2	Longer game states . . . . .	12
8.3	More graphical and audio elements . . . . .	13
<b>9</b>	<b>Conclusion</b>	<b>13</b>
	<b>Appendices</b>	<b>14</b>
<b>A</b>	<b>Development/Experiment Schedule</b>	<b>14</b>
<b>B</b>	<b>Project Budget</b>	<b>14</b>

# 1 Background

In an era of near-ubiquitous computing [6] people are constantly in contact with technology. Moreover, as computerized agents continue to become more sophisticated people have come to rely on technological helpers more over time. Systems like Apple’s *Siri* and Amazon’s *Alexa* are now online in the homes of nearly a quarter of American’s homes according to the Smart Audio Report [1]. Their spring 2020 telephone survey of 1015 US adults revealed a growing acceptance for voice-operated assistants noting that 35% of those surveyed only began using any voice assistant within the year. In parallel, systems capable of offering legitimate physical assistance for an increasing number of tasks are also rapidly coming to market. Tesla’s autopilot is an example of a system with a high degree of authority in the physical world. Their cars have already logged over a billion miles [2] driving on autopilot across thousands of vehicles and the feature is only likely to become more popular (along with its competition from other automakers) in the coming years.

At the surface level, trust is a concept often reserved for the living, for people or maybe our pets. However, at a smaller scale *functional trust* relationships are plentiful, existing between people and the many systems we interact with every day. It doesn’t even have to be a so called “smart” system for these relationships to develop: consider a faulty electric kettle that couldn’t shut itself off, if you can’t trust it to safely stop, you’d probably feel the need to wait and literally watch it boil. These many small functional trust relationships allow us to achieve greater utility from the technologies and tools we use daily by allowing us to delegate parts of otherwise cumbersome tasks to tools that can accomplish them reliably.

A relevant motivational example for this research can be found in Tesla’s new model s redesign coming later in 2021 [8] in which all traces of a gear selector will be removed. According to the company, the new car will be tasked with predicting the drive mode needed given the scenario, either forward or backwards. It’s intriguing to question how often a car would have to get that choice wrong for the user to instinctively override the cars automation by selecting a drive mode on the screen, where they’re still be able to do so. In this example it’s immediately apparent what the computer on-board is doing for the driver. In most cases however, the anticipation offered by already existing software exists behind the curtain of front end user interface (UI). It’s immensely useful however, and important to how we live our lives in 2021 from google maps to product recommendations.

Behind the curtain at Googles I/O keynote this month they announced a project using AI to predictatively group images into patterns based on subject matter a human might not consider interconnected [7]. In practice, certain communities such as people of color have a well founded distrust of digital camera’s ability to accurately capture skin-tone. Although this may be taking it a bit far, its easy to see how little patterns could accidentally reinforce racial bias by grouping different people of color together incorrectly. This example gets at the root of what this research has aimed to understand: how do people react when these systems fail, and how good they have to be for people to truly trust them.

# 2 Related work

Kulms and Kopp [5] conducted a similar experiment attempting to gauge the effects of design anthropomorphism on trust. Their work was partially mirrored, but refocused on anticipation instead. They presented users with a puzzle to complete in cooperation with a computerized agent wherein users could ignore, request and decline, or request and adopt advice from the agent. They varied their independent variable, human-likeness, by switching between three different agents each embodying a specific level of personification: an icon, a CGI rendering of a human, and a recording of an actual person. In order to objectively measure trust they focused on what they call *behavioral trust*, defined as the number of times users exactly followed the agent’s advice throughout the test.

In order to make meaningful assertions about which factors in an interface affect a user’s trust with it, existing formalizations of trust were borrowed from the distributed AI community. Because *trust* is often viewed as a subjective quality of a particular interaction, a definition of trust was needed that could be formalized to only study objective facts about relationships. In their research into trust network design, Josang et al. [4] present a discrimination between two types of trust: *functional trust* and *referral trust* that has been used as a basis for defining trust in the context of this experiment.

- **Referral trust** is trust that someone (an agent) is giving good advice.
- **Functional trust** is trust that someone (an agent) can perform a task.

## 3 Methods and Design

### 3.1 Summary

In this section the goals for the experiment will be outlined. Additionally, the methodology behind this experiment and the design requirements needed for replication will be laid out in detail along with notes on the requirements for the variable compensation scheme, safety, cost, and post-experiment analysis.

### 3.2 Goals

This study aimed to add insight into the trust relationships that users form with computerized agents. Specifically this experiment is focused on examining a hypothesized threshold at which low-quality advice erodes the *functional trust* between users and agents. Using a modified Match-3 game (detailed in section 3.5) reminiscent of popular games such as *Candy Crush Saga* or *Bejeweled* (see *figure 1*), the nature of the *functional trust* that exists between the user and two competing in-game agents was examined.

Across 5 preset scenarios which each allowed the user to play a 4-move modified match-3 game, the strength of two computerized helper agent's simulated anticipation was varied. The key modification made to the core concept of the match-3 format was to restrict the user's choices in game. Instead of allowing any two pieces to be swapped on the board, users were only able to swap one of the two pairs of highlighted icons, one from each agent.

This experiment was conducted with a group of 12 study subjects from the Union College student community (see *figure 7*) remotely via zoom. Subjects were compensated relative to their performance during the game phase of the experiment up to \$10/session. After all 5 rounds, a survey was completed by each subject about their experience. These anecdotal results were utilized in conjunction with logging data from the interface itself about each participant's choices and score in game.

### 3.3 Cost

Any implementation would require a relatively small pool of funding to compensate participants. For the 12 subjects in this experiment, \$111.56 was paid in compensation from a \$300 pool awarded as part of a Student Research Grant from Union College. A detailed breakdown of project budget can be found in appendix B.

#### 3.3.1 Compensation

Compensation is an important part of experiment design intended to form the "carrot on a stick" to motivate participants to make thoughtful choices in-game. In general there should be a reasonable incentive tied to each main round that the participant can earn by doing well. In practice each subject earned \$6 for participation alone with the chance to earn an additional \$1 (up to \$10) for earning above a set amount of points during one of the 4 non-tutorial scenarios indicated at the end of each. A more detailed breakdown of how each subject was compensated can be found in appendix B.

### 3.4 Safety

In accordance with the regulations on student research at Union during the COVID-19 pandemic this experiment was conducted entirely online. As such the risks associated with participation were negligible. Although the experiment presents minimal risks to study subject and experimenter well-being, study subject's emotional safety will be ensured by employing Union HSRC approved informed consent and debriefing media at appropriate times throughout the experiment timeline detailed in appendix A.



Figure 1: *Bejeweled*, a popular match-3 format game currently on the market.



Figure 2: The modified interface displaying two agent's suggestions (outlined in blue and red).

### 3.5 Interface Design

A modified Match-3 game shown in *figure 2* was chosen for its useful characteristics. While multiple different game interfaces are potentially viable for use in this study, they should have a number of distinct qualities in order to provide a useful setting for data collection.

#### 3.5.1 Game Requirements

**High Score Variability** - In order to clearly determine the *relative advice quality* of two agent's suggestions The game interface should have a significant difference between the maximum and minimum number of points possible at each move. The match-3 format offers high points through *cascading moves*, allowing any move to cause a chain reaction scoring hypothetically infinite points.

**Potential for Anticipation** - For the purpose of isolating the effects of an agents anticipatory ability on user perception, a game in which the agents have the potential for a wide range of anticipation is ideal. Anticipation in the context of this experiment is the agent's ability to act on knowledge the user is unaware of. In this case the agents should be able to provide legitimate advice based on future knowledge of game states unknown to the player.

In this experiment, varied levels of simulated anticipation are achieved by directly controlling when cascading moves happen. Since cascading moves incorporate pieces that are unseen to the user, the agent's suggestions are able to provide utility beyond highlighting a viable move the participant can already see. These unknown *future pieces* are revealed after any pieces are cleared from the board below, rendering it impossible to perfectly anticipate future game states without knowledge of those hidden *future pieces*.

#### 3.5.2 Anticipatory Agent

A pair computerized agents capable of selecting viable moves based on current and future game states are critical to the success of this experiment. In this implementation, this anticipation is simulated through a series of pre-determined game states created during the design phase of the interface. In any case, these two agents must be capable of:

**Recognizing viable moves in current game state** - The agents must be able to determine which swaps

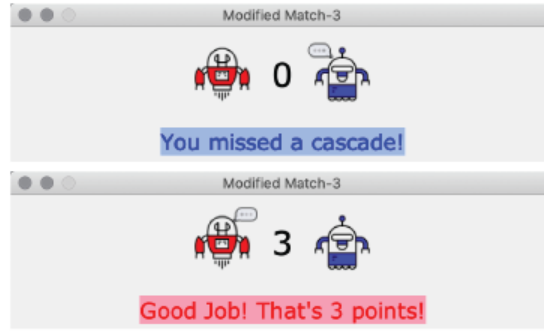


Figure 3: (Top) The blue agent complains that it's move wasn't chosen. (Bottom) The red agent congratulates the user for points scored.

are valid (i.e. which pairs of icons can be swapped to form matches of 3 or more icons) using knowledge of the current state of the game at any point.

**Anticipation based on future pieces** - The agents should be given knowledge of a variable amount of *future pieces* 'above' the play area. Agents should be able to incorporate this knowledge into their predictions to provide the highest possible point output at each turn through their suggestions.

**Notification of the user** - The agents must be able to notify the user when any of three conditions are met (see *figure 3*):

1. If the user chooses an agent's suggestion, then that agent should clearly indicate which icons were successfully matched.
2. If the agent's chosen move causes a cascade (a move with more than one match), then that agent should clearly indicate the amount of cascading matches, as well as the matched pieces in each cascade.
3. If an not-chosen agent's suggestion would have caused at least one more cascade than the chosen agent's, then the not-chosen agent should notify the user of the missed opportunity.

In order to insure consistency across all experiment sessions, the functionality of the anticipatory agents was simulated using preset scenarios.

### 3.5.3 Scenarios

In each experiment session (one per subject) at least two scenarios should be tested. Each scenario should present the user with a different combination game state and anticipatory agent strength. It's critical that the experimenter retains the ability to finely tune the quality of the both the starting arrangement of relevant game components as well as the agent's exact behavior.

In match-3 this can be achieved by laying out the pieces thoughtfully into set scenarios that are consistent across all experiment sessions. This entails deciding where each piece is on the board and setting the exact location that each agent will suggest (see *figure 2*) by hand. This allows for complete control over the points awarded for each choice as well as the relative quality of the advice each agent gives. A complete list of all scenarios implemented can be found in section 4.3.

### 3.5.4 Analytical Tools

The interface itself should be configured to log the user's choices. This logging should produce data that encodes a representation of the actual play area at each turn as well as various information about the game state at each point in a play-through. This implementation is set up to log all of the following to a .csv file for each participant:



**Current score** - How many points has the user scored so far?

**Previous choice** - Which agent did the user choose last turn?

**Each agent's move** - What's the point value of each agent's suggestion?

**Notification status** - Did the unchosen agent notify the user?

## 4 Implementation

### 4.1 Summary

The modified match-3 interface outlined in sections 3.5.1 and 3.5.2 was realized in the Java language. Loosely leaning on an existing representation of a board state provided by Ryan Denlinger's git repository [3] the UI was created entirely using the Swing library for Java. This section will outline the implemented interface and scenarios. Additionally, the chosen methods for subject recruitment, online experimenting, as well as experiment protocol will be detailed.


### 4.2 Game interface

The user-facing interface (pictured in *figure 2*) consisted of a large grid of icons representing the play area. Above the action two iconized images of robots were placed to represent the two color coded agents. The score was always shown at the center of the top section, and space was provided between the play area and the agents for their complaints and notifications to display shown in *figure 3*.

### 4.3 Scenario Building



In order to control the flow of the game within the scenarios, a binary tree based decision tree shown visually in *figure 6* was implemented to hold all of the necessary states for each scenario. Each point of decision is followed by two possible options corresponding to the board states resulting from the blue and red agent's suggestions in the previous move. For readability, subsequent states take the name of the state preceding them followed by a '.' character and a binary digit corresponding to which path a particular state is from (e.g. a board resulting from a red choice would have '.1' appended to its title).

The 5 scenarios, each consisting of their own binary trees, were designed using the *scenario builder* software shown in *figure 5*. Each was set up to deliver a different set of agent behavior and score distribution to tease out insights about the hypothesized trust that might form between subjects and agents in game. The *scenario builder* allowed for the streamlined creation of game scenarios and the decision tree at the heart of the game.

Scenario 1		
Moves 1 - 4	3 (single)	3 (single)

Scenario 2		
Move 1	6 (cascade)	3 (single)
Move 2-4	3 (single)	6 (cascade)

Scenario 3		
Moves 1 - 4	3 (single)	6 (cascade)

Scenario 4		
Move 1	3 (single)	3 (single)
Move 2	7 (cascade)	3 (single)
Moves 3-4	3 (single)	4 (single)



Scenario 5		
Move 1	9 (cascade)	9 (cascade)
Move 2-4	3 (single)	3 (single)

Figure 4: The 5 scenarios used during the experiment, each gray box represents a scenario, with the rows within representing what type of suggestion each agent offered at each move within a given scenario. Note: during scenario 5 the agents were set to deceptively notify on moves 2-4 despite neither agent having a more valuable suggestion.



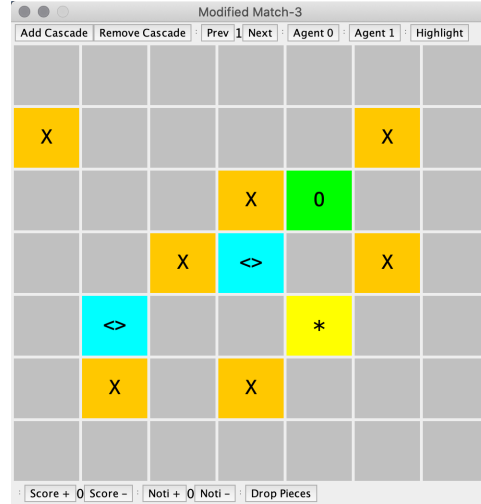


Figure 5: *Scenario builder* Software developed in tandem with the main game to allow for the creation of scenarios. Each box can be clicked to cycle through values, cascading moves can be added (by copying an existing state into a new state) and the point/notification value for each move can be adjusted.

#### 4.3.1 Implemented scenarios

Out of the 5 scenarios, only 4 were useful for actual inquiry. With no point variation between either agent throughout the whole round, the first scenario was instead used as a tutorial to familiarize the participant with the interface and their task. The behavior of all 5 rounds is laid out in the tables shown in *figure 4*.

**Scenario 1:** Tutorial scenario, both agents offer equal moves the entire time.

**Scenario 2:** Presenting the first cascading move, the blue agent offered a 6-point cascade at the beginning of the round followed by subsequent 3-point moves for the remainder. Conversely, the red agent started out by offering a 3-point single move and offered only cascades worth 6 points each for the rest of the round.

**Scenario 3:** Round 3 was set up to give participants a chance to bond with the agents over the course of the game. Since the red agent always had the better move and will complain if the user chooses blue on any move, the prediction was that users would either pick red and stick with red, or pick blue and switch to red eventually.

**Scenario 4:** The 'jackpot' scenario, one agent offered one good move (blue's 7-point cascade in move 2) which is never reinforced. Instead the final two moves present the participant with a better looking 4-point single move from red while blue provides a less-appealing 3-point suggestion.

**Scenario 5:** The final round presented a relatively massive 9-point double cascade from both agents in the first move. For the rest of the game, both agents offered minimum-point moves worth 3-points. Employing deception, the agents in this scenario didn't adhere to their normal notification behavior outlined in section 3.5.2, instead they notified at every move 2-4 despite neither having a more valuable move than its counterpart.

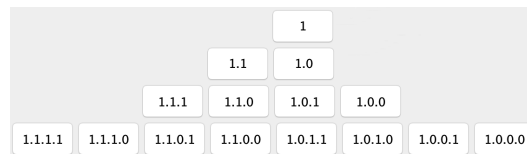


Figure 6: A visual representation of the decision tree used in part of a scenario (moves 1-3).

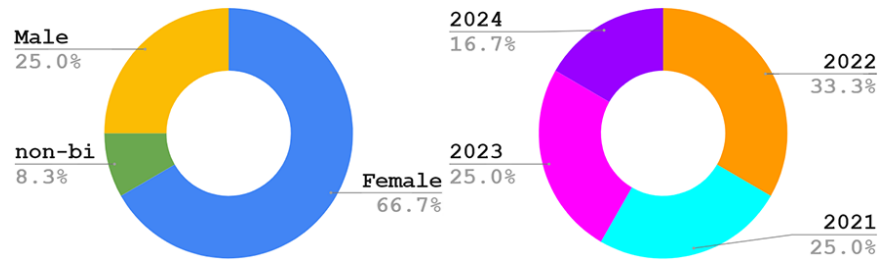


Figure 7: Demographic breakdown of the 12 study subjects.

#### 4.4 Study subjects

As per the current HSRC regulations during the COVID-19 pandemic all study subjects were recruited from inside the Union community. As such, conventional internal methods were used to recruit subjects. All study subjects were recruited using the [survey@union.edu](mailto:survey@union.edu) mass email to reach all students inboxes. The 12 study subjects who responded and participated in experiment sessions were 2/3 female and evenly split across all 4 class years as shown in *figure 7*.

#### 4.5 Online testing venue

Due to COVID-19 restrictions experiments were held via zoom meeting taking advantage of the zoom remote control feature to allow subjects to interact with the interface completely remotely. Venmo and PayPal were used to facilitate payments for the same reason and a google forms document was shared with each participant as the main point of contact responsible for collecting informed consent, demographic information, and later information about their experience as a subject.

#### 4.6 Protocol

Great care was taken to ensure that each experiment session be as identical as reasonably possible between different subjects through careful planning and precisely scripted interactions. The protocol for running a single experiment is as follows:

1. **Setup:** Each experiment started with setting up the interface and testing venue on a Mac laptop. Zoom, Google Forms, and Google Slides in presenter view were all launched before the study subject was invited into the room.
2. **Informed Consent/Demographics Survey:** A link to an informed consent form and google forms response sheet was sent out via zoom chat to each participant which was filled out and "returned" by uploading it to the first section of the google forms survey. Immediately after, participants filled out a short demographics survey using the google form.
3. **Tutorial:** A 1-minute long video outlining how the interface works and the rules of the game was shown to each participant to standardize the tutorial process.
4. **Scenarios:** Participants played the game scenarios 1 - 5 without intervention from the experimenter, a process that took roughly 10 minutes to complete.
5. **Experience Survey:** After all scenarios were completed, the participants returned to the previously distributed google form and completed a short survey about their experience and motivations throughout the game.
6. **Partial Debriefing:** A partial debriefing document was given to each participant as a digital download before exiting the zoom call.
7. **Full Debriefing:** Subjects received a full debriefing during Union's finals week spring 2021 as an email with attached media: a short .pdf debriefing document and a 10 minute presentation on my experiment recorded as a deliverable for this thesis project.

## 5 Results

### 5.1 Summary

In this section the events of the experiment session will be discussed and distilled into conjectural conclusions. First, the insights found in each scenario across all participants will be discussed. Later, the effectiveness of the compensation scheme (see section 3.3.1) in practice as well as a discussion on the participant's ability to make informed decisions based on past knowledge will be conducted.

### 5.2 Observations by scenario

This section will outline the observations made during each scenario across all 12 subjects. Details on the nature of each scenario can be found in section 4.3.1. Due to the small sample size, much of the findings herein are anecdotal, but hint at interesting and non-trivial interactions taking place between the user and agent in game. The concept of *participant retention*, when a participant chooses one agent multiple moves in a row, will be employed throughout the discussion.

#### 5.2.1 Scenario 2

This round offered one of the strangest results out of the entire experiment. It was expected that red choosers would swap to blue after the first round which was backed up by all 5 red choosers swapping. Unexpectedly however, 5 out of the 7 blue choosers swapped to red after the first move. While it might have been easy to explain the red chooser behavior, the blue choosers muddy the water, suggesting that the swap in the second move may in fact be arbitrary.

**5/5 move 1 red choosers swapped to blue after first move:** This initial result was promising as all red choosers managed to find the better agent after one complaint. This was a preliminary indicator that complaints were successful at affecting user behavior, but this notion was immediately contradicted by the move 1 blue choosers.

**5/7 move 1 blue choosers swapped to red after first move:** These subjects were immensely confusing, despite receiving no complaint from the red agent, most chose to switch to red, the better suggester, after the first move. If both groups of participants landed on the same agent these results might have been more powerful.

One explanation for this strange series of events is that the red agent simply had a better looking suggestion on the second move, but this logic falls apart when considering the 5 move 1 red choosers who failed to find the better suggestion move 2. In effect, these results are inconclusive, but they hint at one consistency across all scenarios: subjects just aren't that good at making informed decisions given past information (outlined in section 5.4).

#### 5.2.2 Scenario 3

This scenario also produced surprising results. Here, a clearly superior agent (red) was pitted against blue who gave the worst possible suggestion in the game at each opportunity (a 3 point move). Here we can see some *participant retention* at play, although the behavior of all participants taken together seems to nullify any strong conclusions as a large fraction of participants chose the sub-optimal agent (blue) despite complaints telling them to switch

**Move 1 blue choosers didn't reliably swap to blue given complaints:** Only 1/5 blue choosers chose to swap to red after the first complaint on move 1. The other 4 were split, with 2 swapping after the second complaint on move 2 and the final 2 remaining with blue the entire round despite it being the less viable agent.

**All move 1 red choosers remained with red for at least one additional move:** 4/7 red choosers remained with red the entire game, with the other 3 swapping to blue during the final move. This indicates that a good move associated with no complaint from the opposing agent might cause legitimate *participant retention* in this case. Although three participants eventually choose to try blue in the last round, this can be anecdotally explained by considering that by the last move, blue's suggestion has been displayed for a relatively long period of time and may be enticing purely on the basis of participant curiosity. Another viable explanation for this behavior is more simple: the arrangement of pieces may have looked more interesting to the user, perhaps because a 3-point single move and a 6-point cascade look similar due to the nature of cascading moves requiring unknown pieces to fall.

These were particularly surprising observations given this relatively simple scenario. The hope was that scenario 3 would help determine if the nudge provided by complements and complaints was effective at prompting a swap in agent choice. To that end this scenario was inconclusive, as even after three complaints, a relatively high number of participants failed to switch to the better agent. However, since a 7 red chooser stayed with red for at least one move, and many for the entire game, there was a loose indication of *functional trust* at play within the scenario.

### 5.2.3 Scenario 4

In this so called "jackpot scenario" the strongest evidence for *functional trust* can be found. Here, since there was only ever one notification it was predicted that there wouldn't be a strong enough connection formed between the agent and user after move 2 to prompt decisions based on trust. Surprisingly, and despite red's better looking move in rounds 3 and 4, participants who chose blue exhibited exciting signs of *participant retention* in those final rounds.

**Move 2 red choosers reliably swapped given complaints:** 5/7 participants chose to swap to red after move 2, this indicates that despite a better looking offering from the red agent in the final moves there may have been some *functional trust* formed during move 2, otherwise it would seem much more strange to observe these swaps.

**Most participants chose red in the final round:** Aside from those who chose blue in round 2, all 8/12 remaining participants chose red during the final rounds. This indicated that that 4-point single move, placed as a sort of bait, was effective at its task.

**All 4 move 2 blue choosers remained with blue the entire round:** This was a particularly interesting occurrence, as this contradicts the effectiveness of the bait move suggested by red in the final moves. Although it's impossible to draw concrete conclusions at this point, this behavior indicated possible *participant retention* based on some *functional trust* formed during move 2 due to the blue agent's 7-point cascading move.

In summary, scenario 4 provided a surprisingly consistent pattern of behavior, albeit with a small number of subjects, indicating the possibility, and effects of a *functional trust* relationship in game. The bait move in particular demonstrates how a better-looking suggestion might in fact be dismissed in favor of a move offered from a trusted agent.

### 5.2.4 Scenario 5

Due to the extreme differences in points between the first move and the final 3, it was expected that if *functional trust* was at play, then subjects would remain with their first choice despite being hit with complaints from the other agent. In practice this is exactly what was observed, with a large portion of the group remaining with their first choice for the whole scenario.

**6/12 participants remained with their initial choice the entire round:** Despite the deceptive complaints from the opposing agent, half of all participants remained with their initially chosen agent for all 4 moves.

**Agent complaints did not reliably lead to a swap:** Across all participants, only 42% of complaints were followed by a swap to another agent.

All in all, this scenario provides interesting conjectural evidence of *participant retention* as even among those who didn't remain with their first choice the entire time, participants didn't seem to heed the complaints offered, hinting that they might have been affected by some small level of trust from that first move.

### 5.3 Effect of withholding payments

Withholding payments had a large impact on the mood of each experiment session. Anecdotally, participants who lost points on the second round spend much more time to complete subsequent rounds, perhaps due to their fear of losing more money than they already have. Conversely, the three participants who received the entire compensation were the least stressed, seemingly not as worried about the consequences of under-performing.

One individual who earned the whole \$10 payment was observed "speed running" the experiment, seemingly oblivious to the nudges put in place to guide their decisions, in the end their nonchalant attitude seemed to have worked out in the participants favor. The collected choice data also indicates that those who missed out on compensation were slightly more likely to swap agents when prompted. This all indicates that withholding payments was at least partially successful in promoting thoughtfulness in the study subjects.

### 5.4 User response to past knowledge

Despite the focus of this experiment being on trust relationships there seems to be additional insight to glean about how users interpret and leverage past knowledge in general in the data gathered. Of course this only applies within the bounds of this experiment, but it's interesting to consider how people respond to past knowledge nevertheless, particularly when applying the findings of this research to the real world is the goal.

Taking all 4 main scenarios into account it's hard to find clear evidence that notifications (either complaints or complements) directly impacted subject behavior. However, what this does show is that despite this interface's slow pace, designed to give participants enough time to process the implications of each choice, participants just don't seem that proficient at utilizing that knowledge to inform their actions moving forward.

Scenario 3 (see section 5.2.2) highlights this well. Participants who chose the 3-point suggestion from blue round 1 were expected to take the notification of that fact into account, but generally failed to do so on move 2 with many sticking with their initial choice. Additionally, despite the possible demonstration of *functional trust* in scenario 5 (see section 5.2.4) there is also an indication of poor response to past knowledge in the 6 participants who didn't swap at all, despite the notifications on each turn from the non-chosen agent.

## 6 Validity threats

### 6.1 Inadequate notifications

A major pillar of this experiment was the effectiveness of the notifications offered by the two agents. The concern arises from the relatively small size of the notifications presented by the agents (shown in *figure 3*). Although the notifications are shown during a point in the game where the user is unable to register input, it's still quite small, and can be potentially ignored or even missed by accident. Given that the results indicate that at least some of the notifications were followed, this threat seems partially mitigated by the existing design. However, a change to the interface that more clearly displayed the notification in some way would definitely improve the viability of the experiment. Possible redesigns might include a confirmation box that forced the user to click to confirm each notification (or perhaps just the complaints).

## 6.2 Insufficient graphical and audio elements

Due to the chosen implementation method, using Java Swing UI components, the final interface is far from flashy. Additionally, due to how game states and the decision tree control mechanism (see section 4.3) was implemented, there were no animations between game states. Instead the game played as a sort of slide show, displaying each frame of the game as a single image. There were 3 main elements to displaying each move in a scenario:

- The initial state displaying the agent's suggestions (shown in *figure 2*).
- An intermediate state displaying the matched pieces outlined in green.
- An intermediate state displaying empty space where matched icons disappeared.

After all three frames were displayed the next initial state in the decision tree corresponding to the chosen agent is shown and the above process repeats. The problem with this is that there is no clear graphical indication *in motion* to indicate how game pieces are moving around the board.

This leads to a validity threat exposed during a presentation of preliminary findings to the Union CS department: without clear indications to alert the user to motion and other events in-game, it's reasonable to believe that some subjects may not have fully understood the action taking place at every moment. This is a serious concern for the validity of the experiment, but in practice subjects were able to complete the game with no intervention from the experimenter, suggesting that the interface was ultimately clear enough to convey what it was intended to.

## 7 Ethical concerns

### 7.1 Intentional choice manipulation in the real world

With the insights gained herein, it's clear the humans are at least somewhat receptive to choice manipulation by a computer. This comes at no surprise when considering the absolutely staggering advertising budgets many large companies utilize for targeted advertising. As with any research into human behavior, particularly concerning behavior manipulation, it's easy to imagine the malicious actors leveraging new insights into behavior to try and manipulate people for personal or corporate gain. In the case of this study the findings are likely too inconclusive to provide dangerous information to such actors, but these concerns are legitimate nonetheless and must be thoughtfully considered when embarking on this type of research.

## 8 Future work

### 8.1 Larger sample size

The main limitation of the findings found within this report is the lower than anticipated number of study subjects. Any future work concerning the trust relationships studied here would benefit from a much larger sample size. Ideally, a sample of at least 30 participants would much better serve the research and provide a more solid foundation for drawing meaningful conclusions.

### 8.2 Longer game states

For consistency across experiment sessions each game scenario was entirely pre-decided in the design phase (see section 4.3.1). As such, each consisted of a binary tree of game states which were entirely hand arranged. Future iterations on this experiment design would benefit from deeper binary trees containing more moves. With longer games there is undoubtedly more time for trust relationships to develop, but would require a significantly larger number of scenarios, and much more time.

An additional concern that would hopefully be addressed in a future implementation of this design is the lack of symmetry between the two agent's moves at each stage of the game. Although the scenarios were often designed to present similar value suggestions, no two can truly be equal in a game where each

scenario is a permuted version of the previous as it is in the match-3 format. In the future, more symmetrical boards, or even a different game format could help keep the agent's moves more equal.

### **8.3 More graphical and audio elements**

Following the discussion in section 6.2 any future implementation would benefit from clearer audio/visual markers of action in the game. This could be accomplished in a number of ways including, but not limited to, animations, explosion noises, chimes, or flashing the screen.

## **9 Conclusion**

This experiment was a glimpse at human-computer relationships through the lens of a real world experiment that called on 12 subjects and simplified computerized agents to gain insight into the hypothesized functional trust that may form between the two entities. In doing so a small window into the trust relationships that exist in the real world was opened revealing much, but concluding little, about the complexities of human-computer interactions. Although the sample size and spread results lead to weak and overall conjectural conclusions about the hypothesis presented earlier, it exposed how much more there is to know about how people react to our increasingly anticipatory environment.

The results from the 4 main scenarios indicate that given enough incentive to believe that an agent is offering good advice, people may actually prefer that agent's suggestions in the future over another agent who offers apparently superior advice. Additionally, the effects of withholding payment from the subjects as incentive to thoughtfully participate was assessed and generally seemed successful in its goal. Finally, the general nature of participant's ability to act on past knowledge was called into question, with results suggesting that participants were generally not very proficient in making informed choices based on the past.



# Appendices

## A Development/Experiment Schedule

### Week 1 (March 29th)

- Created scenario concepts

### Week 2 (April 5th)

- Began implementing *scenario builder* back-end software

### Week 3 (April 12th)

- Finished first working prototype of game and *scenario builder* programs

### Week 4 (April 19th)

- Game development
- Scenario debugging week

### Week 5 (April 26th)

- Finalized final experiment scenarios
- Began recruiting study subjects
- Finalized experiment procedure

### Week 6 (May 3rd)

- Began experiment sessions
- Finished recruiting subjects

### Week 7 (May 10th)

- Finished all experiment sessions

### Week 8 (May 17th) SRG spending deadline May 14th

- Drafted poster and presentation

### Week 9 (May 24th)

- Analysis of data
- Finished poster and presentation

### Week 10 (May 31st)

- Work on final thesis report

## B Project Budget

This experiment was solely funded through Union College via a Student Research Grant awarded on March 10th, 2021. The funds total \$300 and were used to provide compensation for 12 single-subject 30 minute experiment sessions. Due to the lower than expected turnout for the experiment, only \$111.56 was payed out to subjects across all sessions.

Expense	Amount
Payment of study subjects	\$300
<i>Study subject rate</i>	<i>\$10/hr</i>
Total projected project cost	\$300
<b>Total actual project cost</b>	<b>\$111.56</b>

Figure 8: The total budget compared with the projected budget of the experiment.

Subject #	Base	Round 2	Round 3	Round 4	Round 5	Total
1	\$6	\$0	\$0	\$1	\$1	\$8
2	\$6	\$1	\$0	\$1	\$1	\$9
3	\$6	\$1	\$1	\$0	\$1	\$9
4	\$6	\$1	\$1	\$0	\$1	\$9
5	\$6	\$1	\$1	\$1	\$1	\$10
6	\$6	\$1	\$0	\$0	\$1	\$8
7	\$6	\$1	\$1	\$1	\$1	\$10
8	\$6	\$0	\$1	\$1	\$1	\$9
9	\$6	\$1	\$1	\$0	\$1	\$9
10	\$6	\$1	\$0	\$0	\$1	\$8
11	\$6	\$1	\$1	\$1	\$1	\$10
12	\$6	\$0	\$1	\$1	\$1	\$9

Figure 9: A compensation breakdown, by subject, of compensation earned per round.

## References

- [1] Apr. 2020. URL: <https://www.nationalpublicmedia.com/insights/reports/smart-audio-report/#download>.
- [2] Jan. 2021. URL: <https://www.tesla.com/VehicleSafetyReport>.
- [3] Ryan Denlinger. *ryanden2018/Match3*. July 2019. URL: <https://github.com/ryanden2018/Match3>.
- [4] Audun Josang, Ross Hayward, and Simon Pope. “Trust Network Analysis with Subjective Logic”. In: *Proceedings of the 29th Australasian Computer Science Conference - Volume 48*. ACSC ’06. Hobart, Australia: Australian Computer Society, Inc., 2006, pp. 85–94. ISBN: 1920682309.
- [5] Philipp Kulms and Stefan Kopp. “More Human-Likeness, More Trust? The Effect of Anthropomorphism on Self-Reported and Behavioral Trust in Continued and Interdependent Human-Agent Cooperation”. In: *Proceedings of Mensch und Computer 2019*. Sept. 2019, pp. 31–42. DOI: 10.1145/3340764.3340793.
- [6] Alex Pentland Maja Pantic Anton Nijholt and Thomas S. Huanag. “Human-Centred Intelligent Human-Computer Interaction: how far are we from attaining it?” In: *Int. J. Autonomous and Adaptive Communications Systems* 1 (Aug. 2008), pp. 168–187. DOI: 10.1504/IJAACS.2008.019799.
- [7] Ian Morris. *Google Photos new upgrades - what are Cinematic moments and Little Patterns?* May 2021. URL: <https://www.tomsguide.com/news/google-photos-upgrades-what-are-cinematic-moments-and-little-patterns>.
- [8] Sean O’Kane. *Tesla’s new Model S will automatically shift between park, reverse, and drive*. Jan. 2021. URL: <https://www.theverge.com/2021/1/29/22256504/teslas-model-s-x-redesign-automatic-shifting-prnd-gears>.