How Does Humanizing Virtual Assistants Affect the Propensity to Follow Their Advice?*

Nanyin Yang [†1], Marco A. Palma[‡1], and Andreas C. Drichoutis[§2]

[1]Texas A&M University

[2]Agricultural University of Athens

Abstract

We investigate the influence of Virtual Assistants' (VA) gender attributes on the propensity of users to delegate an information search task and the potential spillover effects of VA use on subsequent interactions with humans. We conducted two online authority-delegation experiments, varying gender and quality features of human and VA. The results of Study 1 indicate that for high-quality assistants, revealing gender attributes increases delegation to human assistants (HA), but decreases delegation to VA. For low-quality assistants, human males receive higher delegation rates compared to human females and gender-less HA. Strikingly, these results are completely reversed for VA, where delegation rates are higher for female VA and gender-less VA relative to male VA. Study 2 explores the potential spillover effects from interacting with a high/low quality gendered VA on the subsequent decision to select either a male or female HA. The results show only mild spillovers where interactions with low-quality male VA decrease subsequent selection of male HA. Notably, the results of both studies are primarily driven by female respondents. Overall, our findings highlight the impact of assigning anthropomorphic gender features to VA on the propensity to delegate a task with economic benefits, which does not necessarily lead to optimal outcomes.

Keywords: artificial intelligence, authority-delegation game

JEL Codes: C90

[*] Part of this work was completed while Andreas Drichoutis was a Fulbright Visiting Scholar at Texas A&M University.

[†] PhD student, Department of Economics, Texas A&M University, College Station, TX 77843 USA, e-mail: yangnanyin@tamu.edu.

[‡] Professor and Director Human Behavior Laboratory, Department of Agricultural Economics, Texas A&M University, College Station, TX 77843 USA, tel:+1-9798455284 e-mail: mapalma@tamu.edu.

[§] Associate Professor, Department of Agricultural Economics & Rural Development, School of Applied Economics and Social Sciences, Agricultural University of Athens, Iera Odos 75, 11855, Greece, e-mail: adrihout@aua.gr.

1 Introduction

Even though artificial intelligence (AI) has been used for decades, recent advances that provide easy (and inexpensive) access to the masses, have generated a frenzy combination of enthusiasm and fear. While AI enthusiasts foresee countless applications with significant efficiency gains, others are more apprehensive about potential job losses and ethical concerns. AI-powered virtual assistants (VA) have become regular household items that significantly enhance convenience by assuming decision-making responsibilities on behalf of users. VAs have appeared in a multiplicity of domains, including entertainment where among others, Amazon's Alexa and Google Assistant offer personalized recommendation to music and movies; in transportation, VAs offer personalized navigation route planning and even assisted vehicle driving. In recent years, VAs are increasingly overtaking pivotal roles in crucial decision-making processes. For example, medical doctors are using AI to assist with complex healthcare diagnosis and treatment decisions (e.g., the IBM Watson AI system). Similarly, investment firms have adopted AI to inform and enhance their financial investment strategies. When designing VAs, it has become common practice for developers and engineers to embed anthropomorphic characteristics. This practice is grounded in the belief that humanizing VAs can make them seem more familiar and increase the connection, engagement, satisfaction and trust with users of these AI-powered tools (De Visser et al., 2016; Waytz, Heafner, & Epley, 2014).

Anthropomorphizing VAs by incorporating gender features, such as female or male voices, names, and appearance, is one of the most important features in humanizing a VA. The widespread application of gender cues on VAs raise compelling inquiries about their impact on human decision-making. Specifically, it remains unclear how the gender attributes assigned to a computerized VA affect their perception and use, which may ultimately lead to different outcomes. The main hypothesis of this paper is that users may perceive the trustworthiness and capability of a gendered VA differently than a gender-less VA. This in turn can significantly affect whether individuals choose to delegate important decision making tasks to the VA, in areas such as medical treatment or retirement planning. The choices made by users concerning the delegation of such decisions can influence the final outcomes and welfare of the users.

Previous literature raise potential concerns for humanizing VAs, such as assigning them a specific gender through naming conventions, which could inadvertently reinforce prevalent gender stereotypes (Weidinger et al., 2022).[1] It is widely accepted that gender features, such as voices, names and appearance, can effectively manipulate the users' perception of the VA's gender (Jung, Waddell, & Sundar, 2016). Previous research has shown that people tend to transfer the gender and racial biases prevalent in human society onto VAs that contain the corresponding gender or racial attributes. This may reinforce and perpetuate existing gender and racial biases (Cave & Dihal, 2020; Hwang, Lee, Oh, & Lee, 2019). One concern arising from this potential spillover is that the positive or negative experiences with a gendered VA may contribute to the formation of favorable or unfavorable stereotypes toward specific genders. This in turn may negatively impact human to human interactions and foster biases in the way humans engage with each other.

In this paper, we investigate how assigning gender attributes to Virtual Assistants (VA) may impact individuals' willingness to delegate a decision-making search task with economic consequences.[2] More specifically our paper has two main objectives. First, we examine how the gender of a VA influences users' delegation choices and whether these effects mirror patterns observed in interactions with human assistants. Second, we investigate whether interactions with gendered VA (of high or low quality) raise concerns regarding reinforcement

[1] Other concerns reflect the fear of robotic technology adoption overtaking human jobs. For example, Gunadi and Ryu (2023) find that in the United States, a 10% increase in robot exposure is linked to a 2.5% increase in college enrollment rates, with long-term effects pointing to a higher likelihood of individuals obtaining a college degree with greater exposure to robots during their school years.

[2] Bauer, von Zahn, and Hinz (2023) explores delegation decisions to an AI as well, albeit they focus on the AI system explainability on a real estate agents' decision making context. Sunstein and Reisch (2023) find that males tend to favor algorithms more than females, and approximately one-third of individuals already hold a preference for either humans or algorithms that remains unchanged by brief information favoring one over the other. In a large scale field study on ridesharing platforms, Liu et al. (2023) find that drivers are less likely to follow recommendations from an algorithm when the recommendation does not align with their past experience at a given location-time unit and when their peers' actions contradict the algorithmic recommendations.

of prevalent gender stereotypes and biases, potentially affecting subsequent interactions with human assistants.

Our paper consists of two studies each addressing one of the main objectives. In Study 1, we adapt the authority-delegation game in Fehr, Herz, and Wilkening (2013) where participants search for information in an investment decision. Participants have to decide whether to conduct an information search task on their own, which is costly, or to delegate the search to a VA, which is cost-free, but produces sub-optimal returns. In this study we adopt a 2×3 between-subject design, varying the type of assistant (Virtual Assistant or Human Assistant) and the gender attributes (male, female, or gender-less). We manipulate the assistants' gender cues by naming the assistant with a male name, a female name, or providing no name and referring to these as Virtual Assistant or Assistant (in the human treatments). Moreover, we vary the quality of the assistants (i.e., their search capability) on a within-subject basis, which has implications for the optimal choice of subjects: subjects should always delegate to the assistant under *high* quality and should always self-search under *low* quality (an analytical proof can be found in Section 2.2). The results from Study 1 suggest that the impact of gender attributes largely depends on whether the assistant is virtual or human. Specifically, conditional on high-quality assistants, assigning gender features to the assistant *increases* delegation when the assistant is human; conversely, assigning gender attributes *decreases* delegation to VA. In the case of low-quality assistants, a notable pattern emerges: human males tend to receive higher delegation rates compared to human females and gender-less human assistants. Strikingly, these results are completely reversed for VA, where delegation rates are notably higher for female VAs and gender-less VAs when compared to male VAs. These effects are primarily observed among female respondents while male subjects are not sensitive to gender queues.

This observation raises an intriguing point. The overwhelming majority of VAs in the market tend to be predominantly female. This prevalence of female VAs may inadvertently contribute to reinforcing gender stereotypes, particularly in roles associated with assistant-type jobs and may help to explain the high delegation rates for female VAs in low-quality conditions. Perhaps more concerning is the fact that gender-less VAs while not perceived to be female, receive delegation rates on par with females VAs. This suggests that the absence of a gender attribute might itself reinforce prevailing stereotypes in low-quality VA markets where people may perceive assistants to be predominantly female.

Study 2 further explores spillover effects from the interaction with a gendered VA to the subsequent choice of a male vs. a female human assistant. Participants first engage in a block of the authority-delegation game with a VA that varies in gender and quality on a between-subjects basis. After having experienced the VA (high-quality or low quality) in the first block of the authority-delegation game, participants choose whether they prefer to engage with a male or a female Human Assistant for one additional block of the authority-delegation game. We find limited evidence of spillovers from interactions with VA to the choices of male vs. female human assistants. Such effects are mainly observed with low-quality, male-featured VAs, where interactions appear to lower the subsequent selection of male human assistants, primarily among female respondents. We find that there are no significant spillover effects among male subjects.

In recent years people are witnessing a growing integration of AI and robotics in aiding significant decision-making in critical domains, such as medical diagnosis and financial investments. There is a growing body of literature delving into users' perception of AI competence in assisting decision-making and their willingness to utilize algorithms. Dietvorst, Simmons, and Massey (2015) have shown the existence of "algorithm aversion", i.e., humans are less tolerant of mistakes made by algorithms and are less willing to collaborate with algorithms compared to human collaborators. More recently, studies have assessed the real-world adoption of algorithms in medical decision-making. For instance, Agarwal, Moehring, Rajpurkar, and Salz (2023) found that the under-utilization of AI's potential in diagnosis widely exists among radiologists due to their belief updating errors, while Baldauf, Fröehlich, and Endl (2020) found a willingness amongst patients to use AI-driven self-diagnosis apps as supplements to professional medical advice. Furthermore, the influence of AI on users' decision-making and performance has gathered increasing attention, with mixed findings regarding AI's role in team collaboration and performance (Dell'Acqua, Kogut, & Perkowski, 2023; Koster et al., 2022).

Building upon existing research, our study attempts to study users' willingness to delegate decisions to a VA and examines how this decision is influenced by the VA's gender attribute. Our study demonstrates that a VA's gender attribute significantly influences users' delegation decisions, particularly when the VA exhibits high competence. Surprisingly, we find that assigning gender attribute reduces users' delegation frequency towards virtual assistants, while these gender attributes of human assistants enhance users' delegation. Additionally, we highlight the variability in user responses, revealing that female users are more responsive to the gender cues of a VA compared to their male counterparts. Given the novelty of this topic, we selected a delegation task with no a priori expectations to be connected to specific gender stereotypes, but it is possible that the type of task may induce differential outcomes. This is a relevant question for future research.

Additionally, our study adds to the existing literature on autonomy and delegation. Previous studies have shown the prevalence of a preference for autonomy, even when delegating decision making to others is objectively optimal (Ertac, Gumren, & Gurdal, 2020; Fehr et al., 2013). However, most of these studies focus on human-to-human interaction, while both our studies extend to human-to-VA interaction. Consistent with prior research, we observe that respondents are inclined to maintain autonomy even when delegating tasks could yield superior outcomes. Interestingly, our study reveals that individuals are more willing to delegate decisions to VAs compared to their human counterparts.

The rest of the paper is structured as follows. Section 2 presents the methods for the authority delegation game and its parameters. In Section 3 we present the study of the impact of Humanization of AI on delegation decisions and in Section 4 we describe the study of how experience with a humanized AI may spillover to interaction with humans assistants.

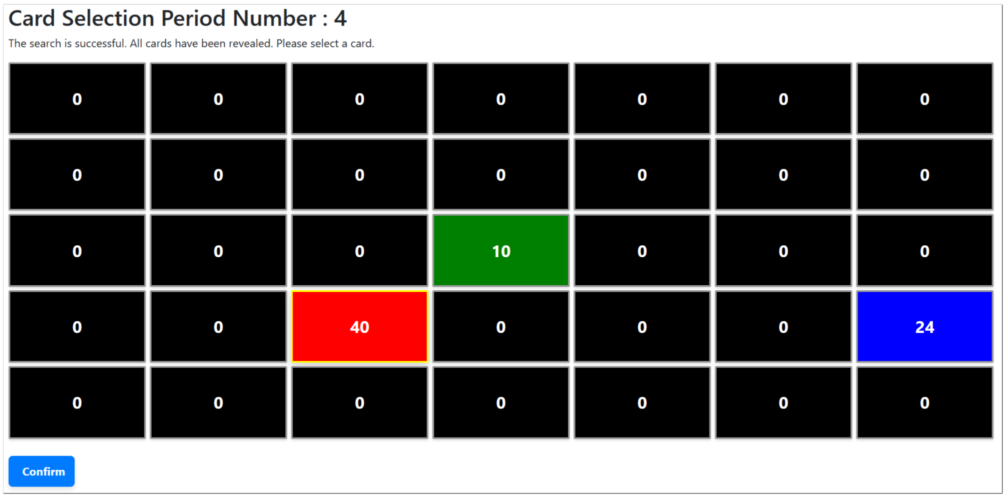
We conclude with discussion and conclusions in Sections 5 and 6, respectively.

2 Methods

We first introduce the structure of the authority-delegation game and explain the choice of parameters in the model and optimal search intensity predictions.[3]

2.1 The Authority-Delegation Game

We adapt the authority-delegation (AD) game in Fehr et al. (2013). In the AD game subjects are presented with a set of 35 cards aligned in 7 columns and 5 rows as shown in Figure 1. There are three cards that offer positive payoffs: the Green Card (returning A_0), the Blue Card (returning A_1), and the Red Card (returning A_2). It is common knowledge that $A_2 > A_1 > A_0$, indicating that the Red Card provides the highest payoff, while the Green Card offers the lowest. The remaining cards are blank and yield no payoff. The cards are shuffled and the Green Card is the only card always visible, while the remaining 34 cards are facing down. The subjects' objective is to select one card that will determine their payoff.



Card Selection Period Number : 4

The search is successful. All cards have been revealed. Please select a card.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	10	0	0	0
0	0	40	0	0	0	24
0	0	0	0	0	0	0

Figure 1: Example Screenshot of A Successful Self-Search

[3] Both our studies were preregistered with the AEA's RCT registry (AEARCTR-0010921) at https://www.socialscienceregistry.org/trials/10921.

Subjects have the opportunity to conduct an information search to reveal all the cards. There are two options available for conducting the information search: "self-search" or "assistant search". The characteristics of the assistants are the manipulating factor we introduce in our treatments described in the experimental design section.

Self-search. Subjects opting to conduct their own information search have to exert costly effort. They specify a costly search intensity $E = \{0, 0.1, 0.2, ..., 0.9, 1\}$, which represents the probability of a successful search. In other words, E represents the probability of subjects discovering the position of all 35 cards which will allow them to select the payoff maximizing card. The search intensity is associated with an effort cost function $g(E) = 25 \times E^2$. If the search is successful, all 35 cards are revealed, allowing the subjects to fully observe the entire set of cards and their corresponding payoffs and maximize their payoff by clicking on the red card. An unsuccessful search reveals only the always visible green card. See Figure 1 for an example screenshot of the revealed card positions after a successful search.

Assistant search. Subjects can choose to delegate the search to an assistant, whose characteristics are predetermined by the treatment conditions as described in the experimental design section. The assistant searches with a fixed intensity of $e = \{e_L, e_H\}$, depending on the quality treatment assignment (low or high) and upon completing the search, the assistant always selects the medium payoff Blue card on behalf of the subject. Note that while the assistant's search is cost-free, a successful search always results in the assistant choosing the medium payoff Blue Card, and not the highest payoff Red Card. If the assistant search is not successful, then the assistant picks the Green Card, thus ensuring participants of at least the lowest payment.

This experimental design introduces a scenario wherein subjects face a trade-off between conducting the search on their own or delegating it to an assistant. Self-searching is costly, but it allows subjects the possibility to select the highest-paying Red Card upon a successful search. On the other hand, delegating the search to the assistant is cost-free, but there is an incentive misalignment: if a search is successful, the assistant always selects the medium payoff Blue Card. Our design is representative of environments where users choose between searching for information independently or relying on VAs before making decisions. Relying on virtual assistants reduces the searching cost and produces a reward that could be improved by engaging in a costly self-information search. There are many emerging markets where consumers delegate searching for products and services to assistants, but due to a lack of perfect customization to fit individual users, there is a misalignment between the recommendation algorithms and the user's specific requirements which denies the maximum payoff.[4]

2.2 Experimental Parameters and Predictions

We set up the parameters to generate an optimal information search choice. The Green Card returns $A_0 = 10$ tokens, the Blue Card returns $A_2 = 24$ tokens, and the Red Card returns $A_1 = 40$ tokens. Tokens are exchanged at a rate of 1 token = \$0.10. Based on the parameter setup, there are two predictions.

Prediction 1 *The optimal search intensity for a self-search is* $E^* = 0.6$.

Prediction 1 is derived from the subjects' expected payoff under self-search: $E \times 40 + (1 - E) \times 10 - 25 \times E^2$. By taking the first and second order conditions, we obtain $E^* = 0.6$ as the optimal search intensity for maximizing subject's payoff.

Prediction 2 *Subjects' optimal choice is to delegate the search to the assistant if the assistant's search intensity* $e \geq \frac{9}{14}$*, and to conduct their own search if* $e < \frac{9}{14}$.

The reasoning behind Prediction 2 is as follows. The expected payoff from self-search is $E^* \times 40 + (1 - E^*) \times 10 - 25 \times E^{*2} = 19$. The expected payoff from the assistant search

[4] Busy lifestyles have resulted in emerging automatic selections for music, food, clothing and many other market goods. For example, prominent companies like Amazon use recommendation algorithms to suggest products based on user behavior and preferences. Similarly, for entertainment streaming, platforms like Netflix or Spotify utilize algorithms to curate personalized content recommendations for their subscribers. However, automated selections may not fully encompass individual preferences and specific needs albeit they do shape users' choices and influence consumption patterns.

with intensity e is $e \times 24 + (1 - e) \times 10 = 14e + 10$. Therefore, the assistant search is more profitable if $e > \frac{9}{14}$, and self-search is more profitable if $e < \frac{9}{14}$. Hence the optimal strategy for participants is to delegate the search to the assistants if their success rate (i.e. quality) is higher than 64% and to self-search otherwise. In our experimental design we vary the quality of the assistant below and above this threshold to generate differences in the optimal search strategy depending on the quality of the assistant. The high-quality treatments have a success rate of 80% and participants should always delegate the search to the assistants, while the low-quality treatments have a success rate of 60% and the participants should always self-search.

3 Study 1: The Impact of the Gender of Virtual Assistants on Delegation Decisions

The goal of Study 1 is to investigate whether assigning a gender to a virtual assistant changes the propensity of participants to delegate the searching decision, given that there exists an optimal delegation strategy based on the structure of the game parameters. The optimal delegation strategy should be unaffected by the gender designation.

3.1 Treatment Conditions

Subjects are randomly assigned to one of six between-subject treatments, varying both the gender characteristics of the assistant and the assistant's type as either a Virtual Assistant (VA thereafter) or a Human Assistant (HA thereafter). Gender was signaled by naming the assistant as either 'Jennifer' or 'Charles' (we discuss the names selection procedures momentarily). For the gender-less treatments we did not provide a name for the assistant:

Female VA. In this treatment the assistant is introduced to the subject as a pre-programmed virtual assistant with humanized female characteristics by naming it "Jennifer".

Male VA. In this treatment the assistant is introduced to the subject as a pre-programmed virtual assistant with humanized male characteristics by naming it "Charles".

Gender-less VA. In this treatment the assistant is introduced to the subject as a pre-programmed virtual assistant, without any reference to gender or name.

Female HA. In this treatment the assistant is introduced as a female Human assistant named "Jennifer", and subjects are informed that the name is a fictitious name chosen by a real human participant who previously recorded the search intensity and the search results.

Male HA. In this treatment the assistant is introduced as a male Human assistant named "Charles", and subjects are informed that the name is a fictitious name chosen by a real human participant who previously recorded the search intensity and the search results.

Gender-less HA. In this treatment, the assistant is introduced to the subject as a participant from a previous session who already chose the search intensity and the search results were recorded without any gender or name reference.

To avoid deception, prior to implementing the main Study, we recruited 13 participants for two in-person sessions and recorded their choices of search intensity and the corresponding search results. The search results were generated from a random-draw program with a given random seed, ensuring that all HAs are identical. We asked participants to choose a pseudonym to represent themselves as human assistants in subsequent sessions. We selected two participants who chose the names "Charles" and "Jennifer" to construct our treatment conditions. We used the computer random-draw generation process to obtain identical responses for the VA and used the exact same names as the human counterparts to represent them. This process results in identical parameters and interactions of participants with the assistants with the only difference being the humanization and gender components associated with each treatment. For more information on how we constructed the assistant treatments through the lab sessions, please check Appendix C in the Electronic Supplementary Material.

For each treatment described above, subjects played two blocks of the authority-delegation game with each block repeating the game for 10 periods. The two blocks differ in the quality of the assistant in terms of the probability of success for conducting the search and were randomized within subjects:

Low-Quality Block. In this block, the assistant's search intensity is $e = 0.6$. According to Prediction 2, subjects' optimal choice is to self-search.

High-Quality Block. In this block, the assistant's search intensity is $e = 0.8$. According to Prediction 2, subjects' optimal choice is to delegate the search to the assistant.

3.2 Experimental Procedures

We implemented an online experiment in oTree (Chen, Schonger, & Wickens, 2016) using general population panelists from Forthright Access, an online research company that handles their own recruitment through a variety of direct advertising channels. All potential panelists are processed through a multi-step, double opt-in procedure to ensure informed consent, and to collect basic profile information. Once participants are in the panel, the company continues to capture new profiling metrics and monitor their data closely. All respondents participate in studies where they are shown monetary compensation in dollar amounts and over half of them have validated their personal phone numbers with the company that allows them to receive instant rewards for their participation.

Subjects first experienced two blocks of the authority-delegation game in random order (10 periods × 2 blocks = 20 periods in total) as described in the previous section. After completing the authority delegation game, subjects participated in a lottery choice task (Dave, Eckel, Johnson, & Rojas, 2010), and completed a brief implicit association test (Sriram & Greenwald, 2009) to capture any individual-level implicit stereotype gender biases. Following these tasks, participants completed a post-experimental questionnaire collecting information about their perceptions of the assistants in terms of their gender characteristics, general trust attitudes towards others, and previous use and perceptions of virtual assistants. A variety of demographic characteristics for the panelists are routinely collected by the company which can be matched to subjects' unique identifier. Subjects were completely anonymous to the experimenters.

We employed several quality controls to account for subjects' attention and comprehension of the experimental procedures. First, all screens that included instructions had minimum threshold timers, and at any point if a subject rushed out of the screen (i.e., less than 3 seconds for short instruction pages or less than 6 seconds for long instruction pages), the subject would be screened out and directed to a final termination screen.[5]

Second, we incorporated three attention check questions.[6] Subjects who rushed through the instructions or failed two or more attention check questions were ejected. A total of 1,472 subjects started the study and 46.2% completed it. In the analysis we use 660 complete responses.[7] See Table 1 for the sample size and descriptive statistics by treatment conditions.

As a manipulation check of our treatments, we asked subjects at the end of the experiment to report their perception of the assistant as a human, female and male and report results in Table A1 in the Electronic Supplementary Material.[8] The average responses from participants show that assigning a gender to the assistant increased their perception as human. Moreover, both the Female VA and Female HA treatments led participants to view the assistant as a female, as intended. Similarly, the Male VA and Male HA treatments yielded a stronger perception of the assistant as male. These outcomes indicate a successful manipulation of the treatment conditions to effectively influence participants' anthropomorphic perceptions of VA while keeping all the quality characteristics constant.

[5] All subjects received a warning at the beginning of the study that they would be excluded at any point if they did not pass a quality control check or did not pay attention.

[6] In the first question, subjects were asked a factual question and were guided to choose the counterfactual answer. Failure to choose the counterfactual answer indicates low attention to the study. At a later point in the study, subjects were explicitly asked to skip the question. In the third question, subjects were asked another factual question and failure to answer it correctly indicates low attention. Any subject that failed two attention checks was excluded from the study and received no payment.

[7] We excluded 20 subjects from a total of 680 valid responses who self-identified as non-binary. This is because we will examine treatment effects depending on subjects' gender. Given the limited representation of non-binary participants, their inclusion in the regression analysis will potentially introduce sample imbalances. To ensure consistency in the analysis, we decided to drop those 20 subjects. Therefore, finally there are 660 subjects included in the analysis. The qualification procedures are acceptable given that we are studying a delegation decision to assistants, which simulates a real-life scenario where users make significant decisions, such as navigation route choices, investment decisions, and health-related decisions. In these scenarios people normally carefully consider whether to rely on an assistant or to make their own decisions. Therefore, it is important to include subjects who read the instructions carefully and pay attention throughout the study.

[8] Screenshots of the questionnaire are included in the Electrinoic Supplementary Material.

Table 1: Descriptive Statistics for Subjects in Study 1

	Treatment Conditions					
	VA Assistants			Human Assistants		
	Female	Male	Gender-less	Female	Male	Gender-less
Female Subjects (%)	45.9	48.7	50.9	48.2	47.3	51.1
Age	45.2	43.3	47.7	45.2	44.8	44.7
	(17.5)	(16.5)	(17.0)	(17.1)	(16.0)	(15.2)
Freq. of Voice Asst. Usage	2.5	2.5	2.4	2.6	2.4	2.6
	(1.5)	(1.5)	(1.4)	(1.5)	(1.4)	(1.4)
Ethnicity (%)						
White	69.2	68.1	74.6	81.5	68.5	82.6
Black/African American	12.0	17.7	13.1	7.4	17.4	6.5
Asian	6.8	4.4	3.5	6.5	7.6	4.4
More than 1 Ethnicity	7.5	5.3	6.1	2.8	3.3	4.4
Native Hawaiian/Pacific	0.8	0.0	0.0	0.0	0.0	0.0
American India/Alaska	0.0	0.9	0.0	0.0	1.1	1.1
Other	2.3	3.5	2.6	1.9	1.1	1.1
Prefer Not To Answer	1.5	0.0	0.0	0.0	1.1	0.0
N	134	115	116	110	93	92

[1] Standard deviations are in parentheses;

[2] “Freq. of Voice Asst. Usage” is the average of responses from the post-experimental questionnaire question “How often do you use any voice assistant, for example Siri, Alexa, or Google?”, and the responses is coded into “1 = I don’t use them or rarely use them, 2 = Once a week, 3 = Once a day, 4 = 2-5 times a day, 5 = More than 5 times a day”.

3.3 Study 1 Results

Figure 2 shows the proportion of delegation to HA and VA by quality of assistant treatment. Recall that regardless of whether the assistant is virtual or human, by design, subjects' optimal strategy is to self-search if the assistant's search intensity is 60% (low-quality assistant), and to delegate the search if the assistant's search intensity is 80% (high-quality assistant). Figure 2 indicates that subjects over-delegate with Low-Quality assistants, and under-delegate with High-Quality assistants. However, subjects delegate significantly more in the High-Quality block compared to the Low-Quality, which is consistent with our model prediction (p-values < 0.001 for signed-rank tests (Wilcoxon, 1945)). More importantly, despite the fact that the assistant type (Virtual or Human) does not affect the expected returns from delegation, subjects delegate significantly more to Virtual assistants than to Human assistants in both the low quality (p-value < 0.001 for a Wilcoxon-Mann-Whitney (WMN) test (Mann & Whitney, 1947; Wilcoxon, 1945)) and the high quality (p-value $=$ 0.006 for a WMN test) treatments. These initial aggregated results seem to point out that participants in general, trust and delegate the search more often to VA rather than HA.

Figure 3 further breaks down subjects' delegation rate by the gender of the assistants. Figure 3a shows the results for the low-quality assistants. The results shows that male VAs receive lower delegation rates compared to female (p-value = 0.018) and gender-less assistants (p-value = 0.019). Interestingly, these results are completely reversed for HAs, with significantly more delegation to male HAs compared to female HAs (p-value < 0.001) and gender-less HAs (p-value < 0.001). These results point out remarkable differences in the perception of the gender of virtual and human assistants. It is worth noting here that most virtual assistants in real life tend to be female and it is possible that the stereotypes of virtual assistants are reversed from human stereotypes due to repeated interactions with female VAs. We explore this issue for the underlying mechanism behind the results in more detail in a later section.

Figure 3b shows the corresponding results for high-quality assistants. The results show

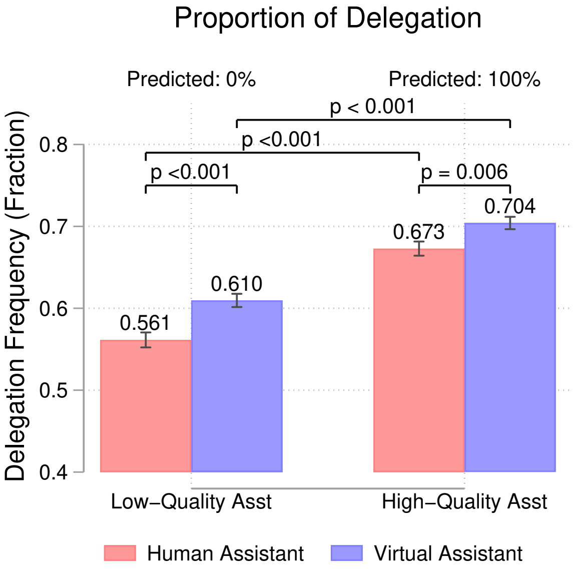
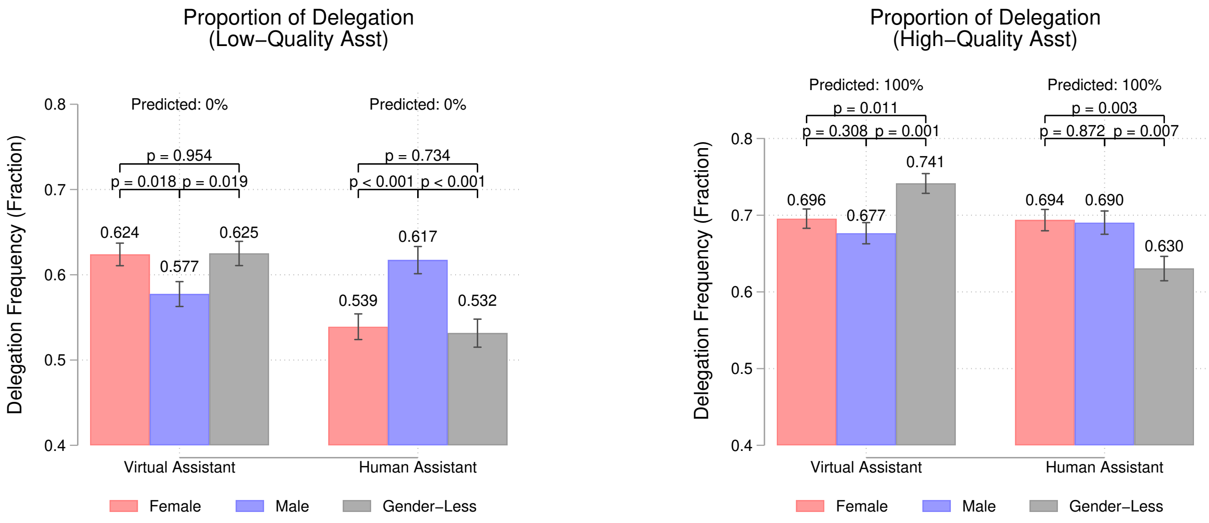


Figure 2: Delegation to Assistants by Assistant's Type and Quality

Note. "Low-Quality Asst" are assistants with search intensity of 60%, and "High-Quality Asst" are assistants with search intensity of 80%. p-values of Human vs. VA assistants are from Mann-Whitney U tests, and p-values of Low- vs. High-Quality assistants are from signed-rank tests.

that the gender-less VA in the high quality condition significantly increases delegation compared to the female (p-value = 0.011) and male (p-value = 0.001) VA; in the case of high quality HA, gender-less HA receive less delegation compared to female HA (p-value = 0.003) and male HA (p-value = 0.007). These findings indicate that the assistants' gender impacts the propensity for delegation and this decision is affected by the quality and type of assistant. In particular, conditional on high-quality assistants, revealing the assistants' gender significantly reduces the delegation to VAs, and it significantly increases delegation to HAs. A possible explanation for this result is that under high quality conditions, participants expect interaction with an expert and the perception of expertise increases when assigning more specific gender features to humans and more impersonal (i.e., computerized) features to virtual assistants. In other words, providing more characteristics for a human expert seems to improve their perception and enhance delegation, while virtual assistants appear to be more useful when presented as a gender-less computer. This finding is consistent with



(a) Delegation to Low-Quality Assistants

(b) Delegation to High-Quality Assistants

Figure 3: Proportion of Delegations by Treatment Conditions

Note. "Low-Quality Assistants" are assistants with search intensity of 60%, and "High-Quality Assistants" are assistants with search intensity of 80%. p-values are from Mann-Whitney U tests.

the concept of the *Uncanny Valley Hypothesis* that predicts that presenting a nonhuman entity with human features enhances its appeal up to a point, but too much human resemblance backfires and results in cold responses (Mori, MacDorman, & Kageki, 2012). Our results show that assigning gender features to a VA strengthens subjects' perception about it as a human, which triggers an effect that aligns with the uncanny valley hypothesis and discourages delegation to the VA.[9]

To verify the robustness of the findings in Figure 3, we run panel logit regressions using a binary indicator variable for delegation to the assistant (vs. self-search) as the dependent variable. We include the following variables as well as all their interactions as independent variables: dummy for a VA (vs. HA); categorical variables for the assistant's gender (female, and male, with gender-less as the base); indicator of assistant's high quality (vs. low-quality); and an indicator for the subject's gender. We also include controls for the periods,

[9] Table A1 in the Electronic Supplementary Material shows that when subjects are asked how they perceive the assistant, they have stronger perceptions of VAs as a human under the female and male conditions compared to the gender-less condition.

subjects' risk preferences, trust in others, age, and ethnicity. Given the complexities arising from the interpretation of the results involving multiple interaction terms, in Figure 4, we illustrate the marginal effects of the gender of assistants on subjects' delegation decisions.[10] In Figure 4a, we illustrate the marginal effects on delegation for female/male vs. gender-less assistants, conditional on the assistant being a HA or a VA, and conditional on the assistant's quality. Notice that in this figure, we aggregate across female and male subjects to calculate the marginal effects. Then in Figures 4b and 4c we further present the marginal effects broken down by the gender of the respondents. The markers in these figures represent point estimates of the marginal effects on the decision to delegate, while the widths of the bars indicate the corresponding confidence intervals at 95% and 90% levels.

As indicated in Figure 4a, when the assistant is a HA, disclosing the assistant's gender — whether male or female— results in a positive marginal effect on the propensity to delegate, regardless of the assistant's quality. Conversely, for VAs, the marginal effects associated with "Female" and "Male" VAs are negative, except for Female VAs in the low-quality condition. Notice that none of these observed marginal effects attain statistical significance, as evidenced by the confidence intervals that cross the zero line.

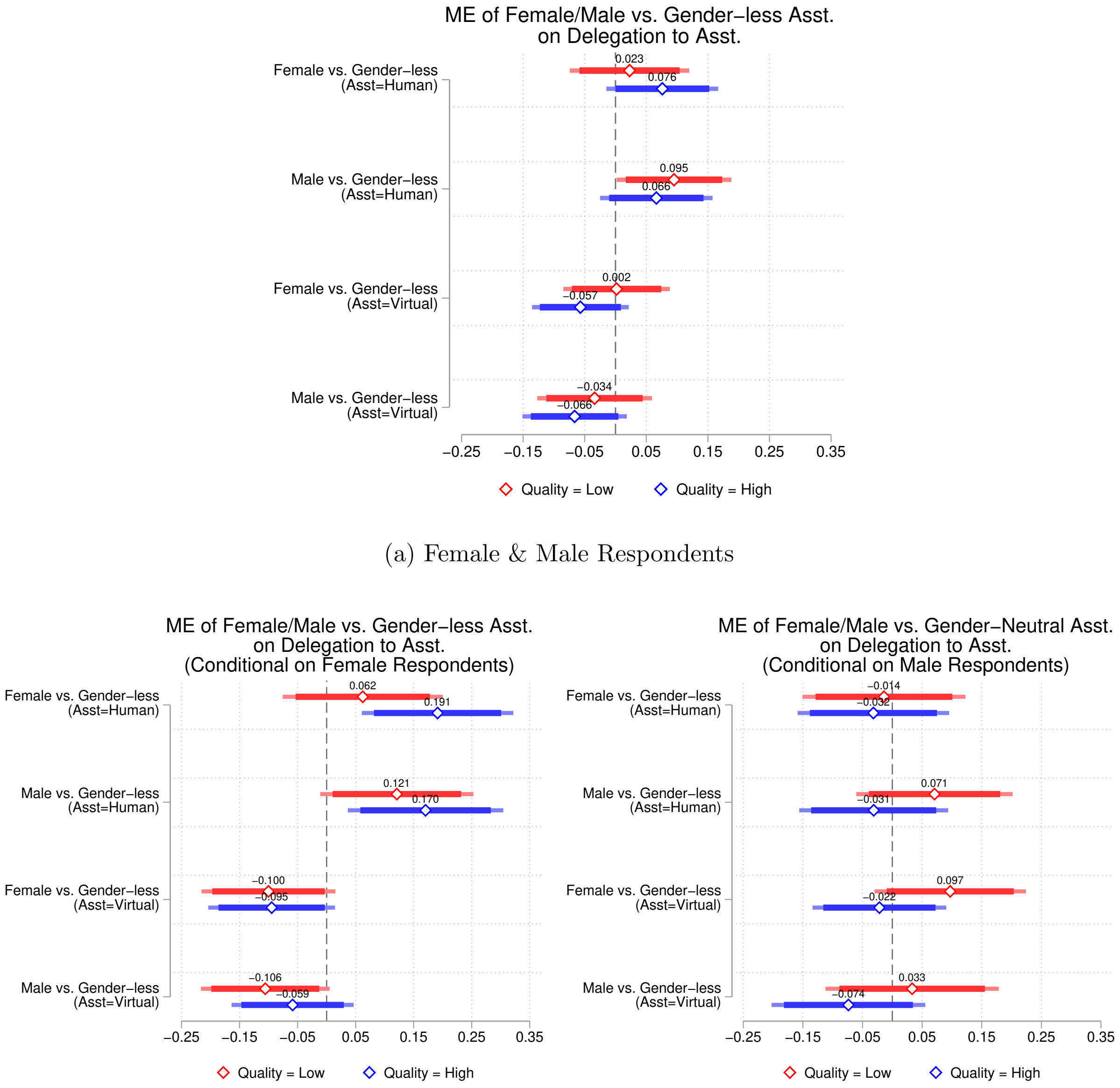
Figure 4b shows that for female respondents, revealing the assistants' gender results in statistically significant impacts on delegation. More specifically, conditional on high-quality HA, both female and male assistants receive significantly more delegation (Marginal effect is 0.191 for Female HA and 0.170 for Male HA) compared to gender-less assistants. Hence, disclosing the gender of HA enhances delegation, particularly when the assistant is of high quality. However, these results are polar opposites when the assistant is not human. In this scenario, revealing the gender of a VA no longer produces an increase in delegation rates. If the VA is of low quality, both the female and male attributes marginally reduce delegation (Marginal effect is -0.1 for Female VA and 0.095 for Male VA).

[10] See Table A2 for the marginal effects from the panel logit regression. We only present Panel A of Table A2 using Figure 4, because the main focus of our study is about the impact of VA assistant's gender on delegation.

Figure 4c shows that for male respondents, the gender of the assistant does not affect their delegation decisions at conventional levels of statistical significance. This null result remains consistent across both VA and HA regardless of the assistant's quality. In essence, while female respondents are sensitive to the gender attributes of the assistant in their delegation decisions, male subjects are unresponsive. This result indicates that female respondents are the drivers of the gender effects observed in the delegation decisions.

Figure 5 depicts the average earnings by treatment. Figure 5a shows the average earnings in the Low-Quality Block where assistants' search intensity is 60%. In this block, if subjects behave optimally (i.e. choose to self-search and search with an intensity of 60%), the expected earnings are \$2.1. Since subjects over-delegate in this condition, their earnings are below this level (p-value < 0.001 for a one-sample t-test of the average payoffs in Low-Quality Blocks being \$2.1). Similarly, in the high-quality condition in Figure 5b, the subjects' average earnings are below the optimal predicted earnings of \$3.4 (p-value < 0.001 for a one-sample t-test of the average payoffs in High-Quality Blocks being \$3.4). This welfare loss is primarily driven by subjects sub-optimal delegation decisions as shown in Figure 3b. Moreover, Prediction 1 posits that the optimal search intensity for a self-search is 60%. However, Figure A1 in the Electronic Supplementary Material highlights that the average search intensities under various treatment conditions do not consistently align with the optimal 60% level under one-sample t-tests. Deviation from the optimal intensity level is particularly pronounced in the High-Quality condition, where subjects tend to under-invest in self-searching. Consequently, this under-investment in self-search further reduces subjects' earnings.

So far, our results have established that the gender of the assistants plays a role in the propensity to delegate. Yet, it remains uncertain whether interactions with gendered VA may impact subsequent perceptions or reinforce gender-related stereotypes prevalent in society. During the experiment, subjects engaged with assistants of two quality levels —low and high quality— in a randomized sequence. While randomization aimed to minimize ordering



(a) Female & Male Respondents

(b) Conditional on Female Respondents

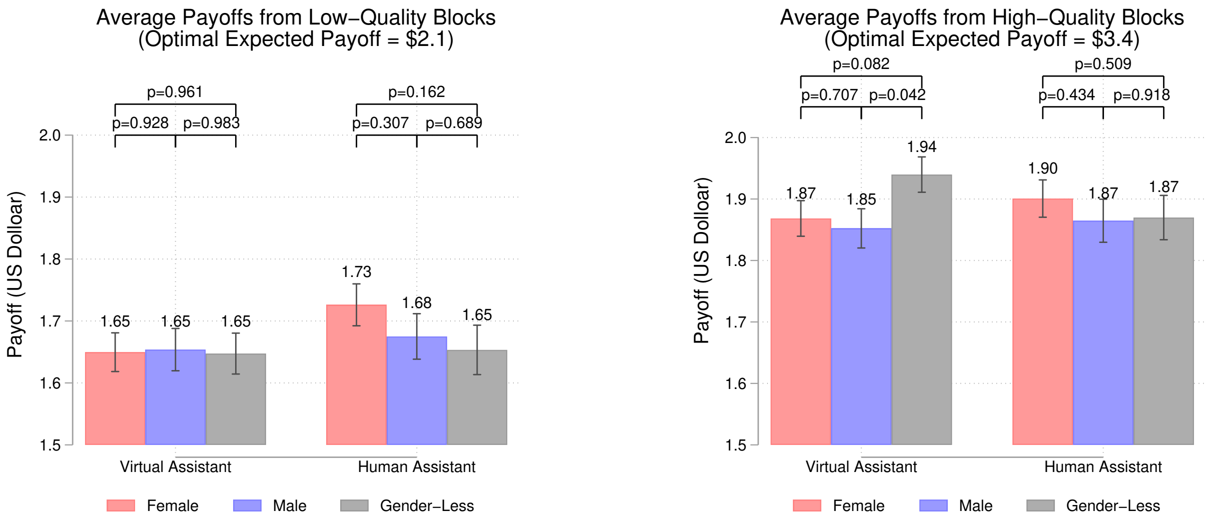
(c) Conditional on Male Respondents

Figure 4: Marginal Effects of Assistants Gender on Respondents' Decisions to Delegate to the Assistant

Note. Marginal effects are from panel logit regression. "Low-Quality Assistants" are periods with assistant search intensity of 60%, and "High-Quality Assistants" are periods with assistant search intensity of 80%.

effects on delegation decisions, it is possible that the sequence of these two quality blocks may have an impact on subjects' preferences for delegation or it may affect gender-related perceptions and stereotypes.

Following the authority-delegation game, subjects responded to a brief Implicit Asso-



(a) Average Payoff in Low-Quality Blocks

(b) Average Payoff in High-Quality Blocks

Figure 5: Average Payoffs by Treatment Conditions

Note: "Low-Quality Blocks" are the experimental blocks (10 periods) where the assistant's search intensity is 60%, and "High-Quality Blocks" are the experimental blocks (10 periods) where the assistant's search intensity is 80%. The "Optimal Expected Payoff" is the expected payoff conditional on the optimal choice in each block, i.e., self-search with intensity 60% in the low-quality block and delegating to assistant in the high-quality block. p-values are from two-sample t-tests.

ciation Test (Brief IAT, from Sriram and Greenwald (2009)). This test measures subjects' implicit associations between gender and pleasant words, and allows us to compute a D-score based on subjects' choices and reaction speeds, where positive values indicate an association between women and pleasant words and negative values indicate an association between men and pleasant words. We created an indicator that equals 1 if the D-score from the test is positive (i.e. the individual has an implicit association between women and pleasant words) and 0 if otherwise (the subject implicitly associates men with pleasant words). Additionally, we created a binary indicator variable "High-Quality First" that equals 1 if subjects encountered the high-quality block before the low-quality block. Results presented in Table 2, show the marginal effect of logit regressions with the positive D-score as the dependent variable on the "High-Quality First" indicator. Columns (1) through (6) show six regression specification results, one for each of the six treatment conditions.

Column (1) of Table 2 is restricted to the Female VA condition, and the results show a statistically significant positive coefficient for the "High-Quality First" indicator. This suggests that, compared to interacting with a low-quality female VA first, the initial interaction with a high-quality female VA increases the probability of a favorable implicit association towards females by 46.2%. In Column (2), for male VAs, the coefficient is negative and significant, also suggesting that encountering a high-quality male VA first reduces the probability of an implicit stereotype favoring females by 46.9% (this can also be interpreted as an increase in the probability of favoring males). In Columns (4) and (5) for the treatments conditions involving HAs, we observe similar patterns. These results highlight the impact of the "first impression". More specifically, interacting with a high-quality assistant first amplifies the respondent's gender preferences towards the corresponding assistant's gender. Interestingly, Columns (3) and (6) demonstrate a different pattern for gender-less assistants. In these cases, experiencing high-quality assistants first favors a positive view towards females, which may not be as a surprising finding if we consider that the majority of VAs in the real world are framed as females.

Insights from Table 2 suggest that the interaction with virtual assistants has discernible effects on subjects' inclinations or stereotypes related to gender. In the next section, we present the results of a second experiment specifically designed to properly identify potential spillover effects from the interaction with virtual assistants in subsequent interactions with humans.

4 Study 2: Do Interactions with Humanized AI Spillover to Human Assistants?

The goal of Study 2 is to investigate the potential spillover effect of the interaction with humanized virtual assistants on subsequent interaction with human assistants.

Table 2: Marginal Effect from Logit Regressions: Spillover of Block Orders to Implicit Association of Gender

	DV: Indicator of D-Score > 0 from Brief Implicit Association Test					
	Virtual Assistant			Human Assistant		
	(1) Female	(2) Male	(3) Gender-less	(4) Female	(5) Male	(6) Gender-less
High-Quality First	0.462***	-0.469***	0.421***	0.447***	-0.457***	0.465***
	(0.016)	(0.059)	(0.045)	(0.048)	(0.027)	(0.024)
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
N of Individuals	127	112	114	108	89	90

[1] Robust standard errors in parentheses; *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$;

[2] Individual controls include gender, age, and ethnicity dummies;

[3] "High-Quality First" is a binary indicator that equals 1 if the subject experiences the block where the assistant with search intensity of 80% first and then the block where the assistant search intensity is 60%.

4.1 Treatment Conditions

To study the spillovers from the interaction with virtual assistants (VA thereafter) to the interaction with human assistants (HA thereafter), we run two blocks of the authority-delegation game, with the first block involving interactions with VA and the second block with HA. Similar to the experiment in Study 1, each block consists of 10 periods.

In Block 1, we manipulate a 3 × 2 between-subjects design involving VA's characteristics: gender (*Female VA, Male VA, Gender-less VA*) and their quality (*Low-Quality, High-Quality*). We use the same names to represent the gender and quality levels as in Study 1. In short, we manipulate the VA's perceived gender by naming it as "Jennifer" (Female VA), "Charles" (Male VA), or just "Virtual Assistant" (Gender-less VA). Consistent with Study 1, for the Low-Quality condition the VA's search intensity is 60%, and for the High-Quality condition this intensity is 80%. Subjects are randomly assigned to one of the six between-subject treatments.

In Block 2, subjects are presented with four HAs, two with female names ("Mary" and "Elizabeth") and two male names ("Thomas" and "Richard").[11] Subjects have to choose

[11] We followed a similar procedure as in Study 1 to construct the HA. We invited participants for an

one of the four available human assistants to interact with throughout the 10 periods of the authority-delegation game in Block 2. Subjects are informed that all HA have a fixed searching intensity level of 80%, i.e. they are all high-quality assistants. As a result, there are no differences among the four HA and the participant's choice reflects only their assistant's gender preference.

In addition, at the end of Block 1 and Block 2, we asked subjects to rate the usefulness and competence of the assistant they had interacted with. The inclusion of these questions helps validate the distinct perceptions participants held regarding the varying quality levels of VA; furthermore, it helps examine whether subjects' perceptions of a VA influenced their perceptions and choices concerning HA. Since the quality of the four HA is the same, we expect to see no differences in the ratings of HA by gender unless there is discrimination in their accreditation.

4.2 Experimental Procedures

We implemented the experiment in oTree (Chen et al., 2016) and collected the data using the Forthright general population panel. Subjects in Study 2 were different than subjects in Study 1. Subjects first experienced two blocks of the authority-delegation game as described in the previous section. Then similar to Study 1, subjects also completed the same lottery-choice task, a brief implicit association test, and a post-experiment questionnaire. We applied the same exclusion criteria as in experiment 1 to ensure valid online responses. A total of 1,279 subjects started the study and the final sample included 637 subjects.[12] See Table 3 for the sample size and descriptive statistics by treatment conditions.

Table B1 in the Electronic Supplementary Material shows the perceptions of VA across treatments for gender as human, male, and female. This table indicates that regardless

in-person study and recorded their choices and the search results using random draws with a specific seed for the search results. Participants chose pseudonyms, and we selected four names from their choices to create the list of names in Study 2.

[12] In addition to unqualified responses, we further excluded 10 participants who self-identified as non-binary from the dataset for same reasons as in Study 1.

Table 3: Descriptive Statistics for Subjects in Study 2

	Treatment Conditions					
	Low-Quality VA			High-Quality VA		
	Female	Male	Gender-less	Female	Male	Gender-less
Female Subjects (%)	38.8	46.4	39.8	45.5	35.9	45.3
Age	47.3	50.0	47.0	50.2	49.0	48.2
	(13.5)	(15.0)	(14.8)	(15.6)	(14.9)	(15.9)
Freq. of Voice Asst. Usage	2.4	2.6	2.6	2.4	2.7	2.7
	(1.5)	(1.5)	(1.4)	(1.5)	(1.5)	(1.5)
Ethnicity (%)						
White	72.6	81.3	72.5	73.0	74.4	74.0
Black/African American	16.7	10.7	11.2	14.0	12.8	9.6
Asian	3.9	2.7	6.1	4.0	4.3	2.9
More than 1 Ethnicity	3.9	3.6	8.2	5.0	6.0	6.7
Native Hawaiian/Pacific	0.0	0.0	0.0	1.0	0.0	1.0
American India/Alaska	0.0	0.0	0.0	0.0	0.9	1.9
Other	2.9	1.8	2.0	2.0	0.9	3.9
Prefer Not To Answer	0.0	0.0	0.0	1.0	0.9	0.0
N	103	112	98	101	117	106

[1] Standard deviations are in parentheses;

[2] "Freq. of Voice Asst. Usage" is the average of responses from the post-experimental questionnaire question "How often do you use any voice assistant, for example Siri, Alexa, or Google?", and the responses is coded into "1 = I don't use them or rarely use them, 2 = Once a week, 3 = Once a day, 4 = 2-5 times a day, 5 = More than 5 times a day".

of the VA's quality, subjects tend to perceive female-named VA as more female than male, and vice versa for male-named VA. This confirms the successful manipulation of participants' perceptions of VA's gender. Moreover, in Table B2 in the Electronic Supplementary Material we report subjects' average perceptions of the gender of HA in Block 2. The results indicate that female-named HAs are predominantly perceived as female, while male-named HA are perceived as male more than female. These findings affirm the consistency of subjects' perceptions of human assistants' gender in Block 2 with our experimental manipulations.

4.3 Study 2 Results

Figure 6a shows the proportion of subjects choosing a female-named HA in Block 2 by treatment. Overall, more than 60% of subjects choose female HA in Block 2 regardless of the random treatment manipulations in Block 1. In the Low-Quality condition, the proportion of subjects opting for a female HA peaks at 69.6% under the Male VA condition. This result shows that after interacting with a low-quality male VA, participants show the highest rates of female HA in block 2. This proportion is higher than under the Female (63.1%) and Gender-less (63.3%) conditions. In the High-Quality condition, female HA are chosen at the highest rate after participants interacted with a female VA (68.3%). However, there are no statistically significant differences from Mann-Whitney U tests among the three gender treatment conditions within each quality level. Therefore, we do not find evidence for spillovers from the interactions with VA to the choices of female vs. male HA. Next, we will investigate this spillover effect further by narrowing respondents by their gender.

When restricting the sample to only female respondents, Figure 6b shows that conditional on Low-Quality VA in Block 1, the proportion of subjects choosing female HA is significantly higher under the Male VA condition (73.1%) than under the Female VA condition (52.5%) and Gender-less VA condition (43.6%). In other words, encountering a low-quality male VA reduces subjects' choices for male HA in the subsequent block. However, for the High-Quality condition, the Female VA treatment does not yield a significantly higher proportion of choices for female HA, despite the proportion being higher than those under the Male VA and Gender-less VA treatments.

Conversely, when restricting the sample to male respondents as depicted in Figure 6c, male subjects' choices of HA's gender in Block 2 seem to have no effect in the choice of gender, given that none of the Mann-Whitney U tests between treatment conditions is statistically significant. This is similar to our findings in Study 1 that female respondents are more responsive than males to the VA's gender.

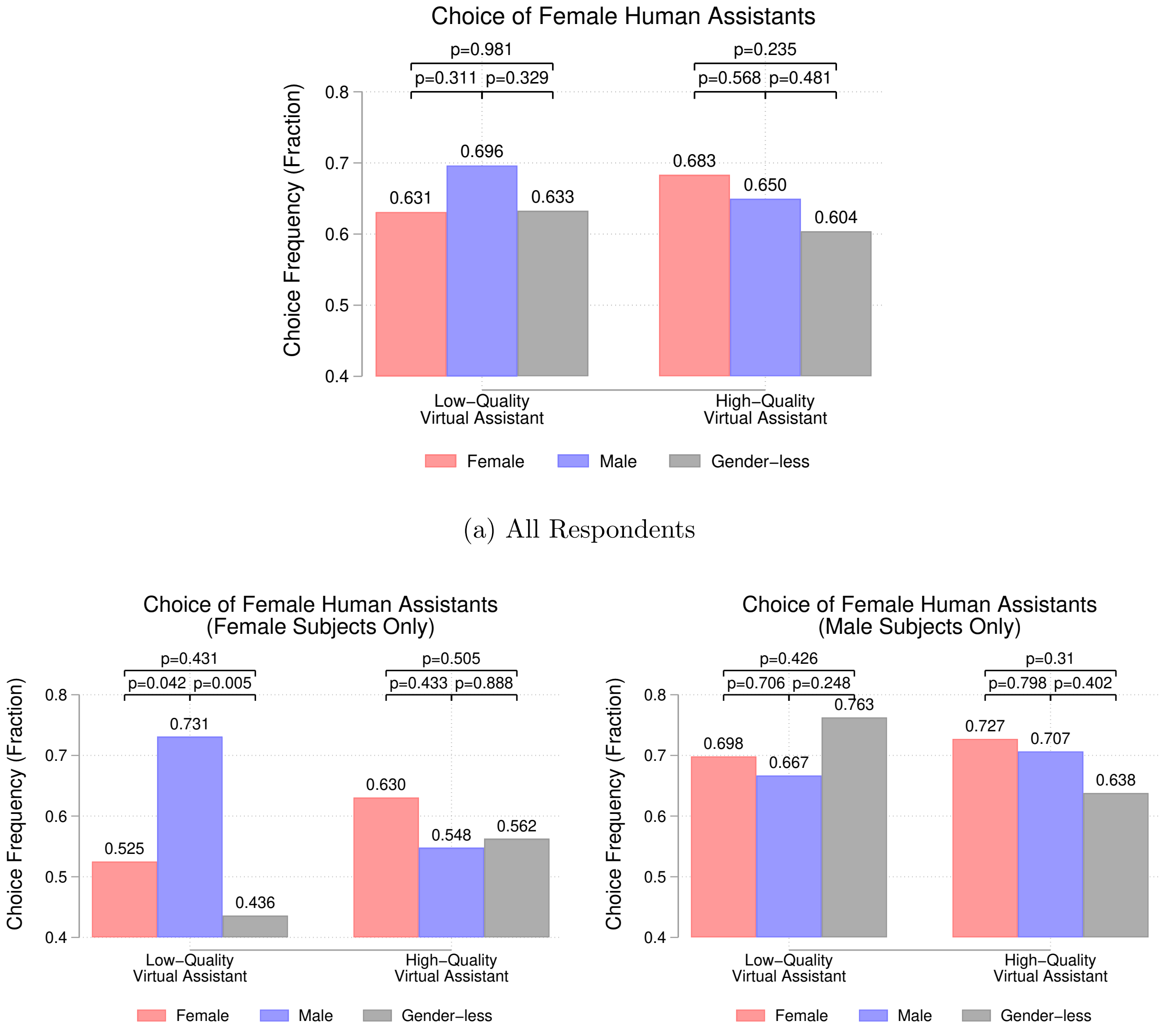
To validate the findings from Figure 6, we conduct a panel logit regression with the indicator of choosing female HA as the dependent variable (vs. choosing a male HA). We include all the combination of the following variables as the independent variables: Indicator of the VA as high quality (vs. low quality); categorical variables for the VA's gender (female and male, with gender-less as the base); indicators of the respondent's gender as female (vs. male). We also include control variables for periods, respondents' risk preferences, trust in others, age, and ethnicity. Table 4 reports the marginal effects of the estimation results. Panel A of this table shows the marginal effects of Female VA and Male VA on the choice of a female HA conditional on the quality of VA. Panel B shows the marginal effects of the High-Quality condition vs. the Low-Quality condition, conditional on VA assistants' gender attributes. In Column (1), we aggregate male and female respondents to compute the average marginal effects. Columns (2) and (3) provide the marginal effect for female and male subjects, respectively.

Panel A reports marginal effects of the gender of the VA (with gender-less being the base) on the probability of selecting a female HA. Column (1) shows that the gender attribute of the VA does not yield statistically significant effects on the choices of female vs. male HA. None of the marginal effects exhibit statistical significance, holding true for both high and low-quality VA. In Column (2), when focusing on female subjects, there is a statistically significant marginal effect associated with the coefficient "Male vs. Gender-less" in the Low-Quality condition (Marginal effect is 0.302). This effect indicates that conditional on the low quality VA, interacting with a Male VA increases the likelihood of subjects opting for a female HA by 30.2%. In other words, interacting with a low-quality male VA significantly reduces subjects from selecting male humans. Finally, in Column (3), concentrating on male subjects, no statistically significant marginal effects arise from VA's gender attributes. According to the findings from Panel A, the spillover effect from VA's gender attributes onto the selection of HA is limited in scope: it primarily diminishes the female subjects' preferences for male HA when they are initially exposed to a low-quality male VA.

Furthermore, Panel B examines the spillover effects of high-quality VA on subjects' choices of HA. Column (1) pools marginal effects for female and male respondents and shows no statistically significant effects of the quality of VA on the probability of a subject choosing a female HA (vs. a male HA). Column (2) concentrates on marginal effects conditional on female respondents, where a high-quality VA significantly reduces the choices of female HA, solely in the context of Male VA (Marginal effect is -0.192). This finding suggests that when interacting with a male VA, the high quality of the VA decreases (increases) female respondents' choices of female (male) HA. In Column (3), centering on male subjects, VA's quality does not yield significant effects on any of the outcomes of interest. These insights underscore that, within our experimental framework, the spillover effect of gendered VA assistant quality on human gender preferences manifests in a very specific circumstance – when females (but not males) are exposed to a high-quality male VA assistant.

5 Discussion

Does assigning gender attributes to virtual assistant matter for human users' willingness to delegate decision-making to the VA? And do these gender cues spillover to human users' perception of other humans of a particular gender? Our experimental procedures control for the quality of the assistants to ensure that there are no differences among them except for the framing of their gender manipulated through the use of names and whether they are human or virtual. The results indicate that the gender attribute attached to a virtual assistant does indeed influence users' delegation decisions, but the impact differs depending on whether the assistant is virtual or human. For low-quality Virtual Assistants, a male attribute reduces delegation by female users to the virtual assistant, yet it encourages delegation to human assistants. Conversely, for high-quality virtual assistants, anthropomorphizing with gender attributes, reduces female users' probability of delegation to the virtual assistant but increases their probability of delegation to human assistants. Interestingly, male users are unresponsive to the gender attributes of either virtual or human assistants. However, the spillover from interacting with virtual assistants to human assistants' gender preference is limited —negative experiences with a male virtual assistant reduce female users' subsequent preference for male human assistants.



(a) All Respondents

(b) Female Respondents

(c) Male Respondents

Figure 6: Choice of Female-Named Human Assistants in Block 2 by Treatments in Block 1

Note: This figure reports the proportion of subjects choosing a female-named human assistant over a male-named assistant in Block 2, by treatment conditions in Block 1. The left panel restricts to female subjects only; the right panel restricts to male subjects only. There are 4 non-binary respondents excluded. *p*-values are from Mann-Whitney U tests.

Table 4: Logit Regression: Marginal Effects of VA's Gender on Choice of Female Human Assistants

	DV: Indicator of Choosing Female Human Asst		
	(1)	(2)	(3)
	Male & Female Subjects	Female Subjects	Male Subjects
Panel A: Female/Male VA vs. Genderless VA (Conditional on VA Quality)			
Female vs. Neutral (Quality=Low)	0.006	0.077	-0.045
	(0.065)	(0.110)	(0.078)
Female vs. Neutral (Quality=High)	0.087	0.093	0.082
	(0.066)	(0.099)	(0.089)
Male vs. Neutral (Quality=Low)	0.070	0.302***	-0.101
	(0.063)	(0.100)	(0.082)
Male vs. Neutral (Quality=High)	0.043	-0.014	0.086
	(0.064)	(0.102)	(0.083)
Panel B: High-Quality vs. Low-Quality VA (Conditional on VA Gender)			
High vs. Low Quality (VA=Female)	0.060	0.141	0.001
	(0.065)	(0.104)	(0.083)
High vs.Low Quality (VA=Male)	-0.047	-0.192**	0.060
	(0.061)	(0.096)	(0.081)
High vs. Low Quality (VA=Gender-Neutral)	-0.020	0.125	-0.127
	(0.066)	(0.106)	(0.085)
Individual Controls	Yes	Yes	Yes
N of Individuals	633	633	633

[1] Robust standard errors in parentheses; $*p < 0.1$, $**p < 0.05$, $***p < 0.01$;

[2] "VA" is the acronym of virtual assistant;

[3] Individual controls include age, ethnicity dummies, subjects' risk tolerance measured by the lottery game from Dave et al. (2010), and subjects' trust in others measured by post-experimental survey questions;

[4] Subject demographic information (gender, age, and ethnicity) was primarily collected and provided by Forthright Access. There were 4 subjects in the dataset whose demographic information was missing, so they are not included in the analysis in this table.

Assigning gender attributes to algorithms and robotics is a widespread practice in the diffusion of technology. In practical contexts, virtual assistants are often predominantly characterized as female, possibly due to the perception that women are more inclined to possess qualities such as friendliness, nurturing, and empathy —aligning with societal stereotypes that are believed to enhance user engagement. Borau, Otterbring, Laporte, and Fosso Wamba (2021), through online experiments, demonstrated a preference for female chatbots over their male counterparts, as they were perceived as more human-like and more considerate. It is also possible that the implied characteristics of female assistants may dangerously spread stereotypes of women as more submissive and obedient assistants.

Nonetheless, the results of Study 1 reveal that the effects of attributing gender to both human and virtual assistants vary in their impact on users' delegation decision. Specifically, when an assistant demonstrates high competence, assigning gender attributes significantly reduces the frequency of users' engagement with virtual assistants, leading to sub-optimal outcomes. In contrast, these same attributes encourage delegation to human assistants. These findings suggest that users perceive gender attributes differently when associated with a virtual assistant compared to a human assistant. Therefore, we caution against relying solely on traditional gender perceptions within human society to forecast the influence of such gender attributes on users' perceptions and interactions with virtual assistants.

Taking a step further, our study unveils the heterogeneous effect of an assistant's gender attributes on male and female users. In our experimental environment, we observed that female participants exhibited increased sensitivity to the gender attributes of both virtual and human assistants, whereas male participants' delegation decisions remained unaffected by this information. This disparity may be attributed to a variety of factors, including societal norms, variations in sensitivity to gender differences, or divergent personal experiences shaped by gender. These complexities warrant further exploration.

Given the observed gender-based differences, we cautiously suggest that the gendering design of a virtual assistant should consider users' gender as a significant variable. For instance, if the predominant user demographic of a virtual assistant leans toward being female, it is reasonable to anticipate that users' engagement and usage patterns could be more influenced by the virtual assistant's gender features. Such considerations should be integral to the design process to enhance user satisfaction and engagement.

An important concern regarding the gendering design of virtual assistants is the potential for these gender attributes to perpetuate or reinforce gender stereotypes (Weidinger et al., 2022). To illustrate, consider our study's setting: interacting with a low-quality virtual assistant of a specific gender may lead to the creation of negative impressions about that gender, subsequently influencing users' attitudes towards individuals of the same gender. In Study 1, we indeed observe evidence of such spillover effects. Participants who had interactions with both high and low-quality virtual assistants, exhibited spillover effects based on their initial impressions —specifically, the quality level in the first block of the delegation task, influenced their gender biases, as measured by a brief implicit association test conducted after the main game. Consequently, positive experiences with a gendered virtual assistant in the first block led to a more favorable disposition towards that gender, while negative experiences yielded the opposite effect.

However, in Study 2, we only observed limited evidence of spillover effects from interactions with virtual assistants to participants' preferences when selecting between male and female human assistants. This suggests that concerns about interactions with virtual assistants significantly altering human users' preferences for human assistants performing similar tasks, may not be strongly justified. Nonetheless, it is crucial to recognize that this does not negate concerns about the potential for gender biases to be perpetuated or reinforced through the anthropomorphism of robotics and algorithms. The divergent findings on spillover effects of gendered virtual assistants between Study 1 and 2 underscore the role of context, task nature, and initial interaction quality in shaping and reshaping individuals' perceptions of gender.

One limitation of the current study is that it does not delve into examining the mechanisms underlying the divergent outcomes of attributing gender characteristics to human versus virtual assistants and their impact on users' delegation decisions. We posit two potential explanations that merit more in-depth investigation. Firstly, participants may hold varying expectations for human and virtual assistants. While they are inclined likely to embrace and value human-like characteristics in human assistants, their priorities might shift towards functionality and competence when assessing virtual assistants. This contrasting emphasis on attributes could play a pivotal role in shaping their delegation choices.

Secondly, as discussed by Mori et al. (2012), anthropomorphizing virtual assistants might trigger a sense of discomfort among users, a phenomenon known as the "uncanny valley" effect. Assigning names to the virtual assistants in our study could have potentially induced some discomfort among participants, diminishing their trust in the virtual assistant. In light of these considerations, it becomes evident that further research is required to uncover the underlying mechanisms contributing to the differential effects observed in the delegation of tasks to humanized virtual and human assistants. A comprehensive understanding of these mechanisms will serve as guidance in designing virtual assistants with the ultimate goal of enhancing user experiences and mitigating unintended loss resulting from anthropomorphism.

6 Conclusion

With the development of artificial intelligence and the recent breakthrough in large language models, AI-powered virtual assistants are increasingly involved in critical decision making processes. Humanizing virtual assistants by incorporating gender attributes is a common practice, with the belief that this enhances user engagement and trust. In this paper, we investigated how gender attributes assigned to a virtual assistant influence users' choices when it comes to delegating their decision-making responsibilities. We find compelling evidence of an impact of these attributes on users' delegation decisions. Moreover, our findings reveal that this impact varies notably when compared to human assistants whose genders are explicitly designated. In particular, when the virtual assistant is of high quality, assigning gender attributes significantly reduces users' delegation, resulting in suboptimal outcomes.

We also delve into the potential spillover effects stemming from interactions with gendered virtual assistants on users' attitudes toward individuals of the corresponding gender. Our results unveil that this spillover only occurs among female users when interacting with low-quality male virtual assistants.

This study contributes substantially to the human-computer interaction literature, emphasizing the importance of pertaining caution when designing the gender features of virtual assistants. It underscores the nuanced differences in the impact of gender cues between human and virtual assistants, thereby cautioning against the uncritical application of findings derived from human-human interactions to the complex realm of humanizing of virtual assistants.

References

Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2023, July). Combining human expertise with artificial intelligence: Experimental evidence from radiology [Working Paper]. (31422). Retrieved from http://www.nber.org/papers/w31422 doi: 10.3386/w31422

Baldauf, M., Fröehlich, P., & Endl, R. (2020). Trust me, i'ma doctor–user perceptions of ai-driven apps for mobile health diagnosis. In *Proceedings of the 19th international conference on mobile and ubiquitous multimedia* (pp. 167–178).

Bauer, K., von Zahn, M., & Hinz, O. (2023). *Please take over: XAI, delegation of authority, and domain knowledge* (SAFE Working Paper No. 394). Leibniz Institute for Financial Research SAFE.

Borau, S., Otterbring, T., Laporte, S., & Fosso Wamba, S. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of ai. *Psychology & Marketing*, *38*(7), 1052–1068.

Cave, S., & Dihal, K. (2020). The whiteness of ai. *Philosophy & Technology*, *33*(4), 685–703.

Chen, D. L., Schonger, M., & Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88-97. Retrieved from https://www.sciencedirect.com/science/article/pii/S2214635016000101 doi: https://doi.org/10.1016/j.jbef.2015.12.001

Dave, C., Eckel, C. C., Johnson, C. A., & Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, *41*, 219–243.

Dell'Acqua, F., Kogut, B., & Perkowski, P. (2023, 04). Super Mario Meets AI: Experimental Effects of Automation and Skills on Team Performance and Coordination. *The Review of Economics and Statistics*, 1-47. Retrieved from https://doi.org/10.1162/rest_a_01328 doi: 10.1162/rest_a_01328

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Eckel, C. C., & Grossman, P. J. (2008). Men, women and risk aversion: Experimental evidence. *Handbook of experimental economics results*, *1*, 1061–1073.

Ertac, S., Gumren, M., & Gurdal, M. Y. (2020). Demand for decision autonomy and the desire to avoid responsibility in risky environments: Experimental evidence. *Journal of Economic Psychology*, *77*, 102200.

Fehr, E., Herz, H., & Wilkening, T. (2013). The lure of authority: Motivation and incentive effects of power. *American Economic Review*, *103*(4), 1325–1359.

Gunadi, C., & Ryu, H. (2023, October). How do people respond when they know that robots will take their jobs? *Oxford Bulletin of Economics and Statistics*, *85*, 939-958.

Hwang, G., Lee, J., Oh, C. Y., & Lee, J. (2019). It sounds like a woman: Exploring gender stereotypes in south korean voice assistants. In *Extended abstracts of the 2019 chi conference on human factors in computing systems* (pp. 1–6).

Jung, E. H., Waddell, T. F., & Sundar, S. S. (2016). Feminizing robots: User responses to gender cues on robot body and screen. In *Proceedings of the 2016 chi conference extended abstracts on human factors in computing systems* (pp. 3107–3113).

Koster, R., Balaguer, J., Tacchetti, A., Weinstein, A., Zhu, T., Hauser, O., ... others (2022). Human-centred mechanism design with democratic ai. *Nature Human Behaviour*, *6*(10), 1398–1407.

Liu, M., Tang, X., Xia, S., Zhang, S., Zhu, Y., & Meng, Q. (2023). Algorithm aversion: Evidence from ridesharing drivers. *Management Science (Ahead of print)*, *0*(0). doi: 10.1287/mnsc.2022.02475

Mann, H. B., & Whitney, D. R. (1947). On a test whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50-60.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, *19*(2), 98–100.

Sriram, N., & Greenwald, A. G. (2009). The brief implicit association test. *Experimental psychology*, *56*(4), 283–294.

Sunstein, C. R., & Reisch, L. (2023, August 18). Do people like algorithms? a research strategy. Retrieved from https://ssrn.com/abstract=4544749 doi: 10.2139/ssrn.4544749

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology*, *52*, 113–117.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., ... others (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 acm conference on fairness, accountability, and transparency* (pp. 214–229).

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80-83.

Electronic Supplementary Material of

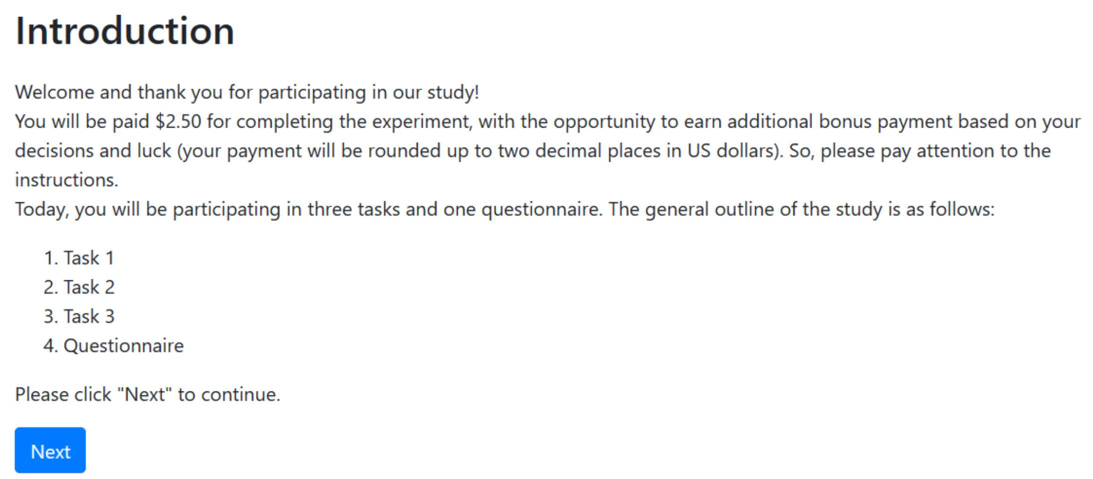
How Does Humanizing Virtual Assistants Affect the Propensity to Follow Their Advice? An Experimental Investigation

Nanyin Yang[*]
Marco A. Palma[†] and
Andreas C. Drichoutis[‡]

Study 1: Experimental instructions

These are screen captures from oTree.

Instructions



Introduction

Welcome and thank you for participating in our study!
You will be paid $2.50 for completing the experiment, with the opportunity to earn additional bonus payment based on your decisions and luck (your payment will be rounded up to two decimal places in US dollars). So, please pay attention to the instructions.
Today, you will be participating in three tasks and one questionnaire. The general outline of the study is as follows:

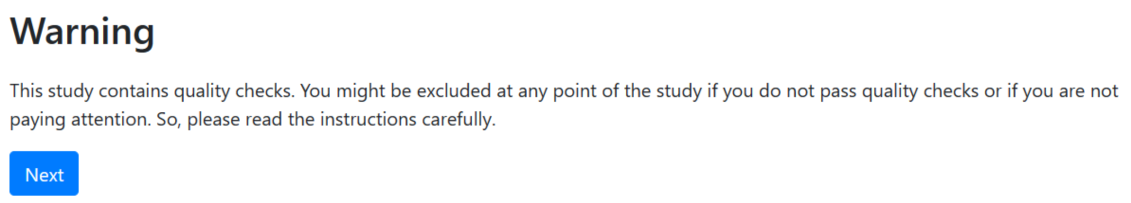
1. Task 1
2. Task 2
3. Task 3
4. Questionnaire

Please click "Next" to continue.

[*] PhD student, Department of Economics, Texas A&M University, College Station, TX 77843 USA, e-mail: yangnanyin@tamu.edu.

[†] Professor and Director Human Behavior Laboratory, Department of Agricultural Economics, Texas A&M University, College Station, TX 77843 USA, tel:+1-9798455284 e-mail: mapalma@tamu.edu.

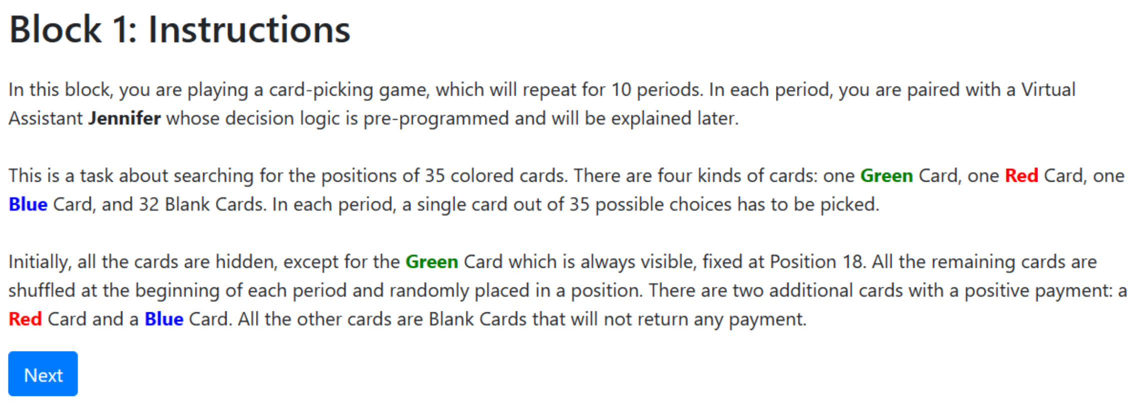
[‡] Associate Professor, Department of Agricultural Economics & Rural Development, Agricultural University of Athens, Iera Odos 75, 11855, Greece, e-mail: adrihout@aua.gr.



Warning

This study contains quality checks. You might be excluded at any point of the study if you do not pass quality checks or if you are not paying attention. So, please read the instructions carefully.

The female VA treatment

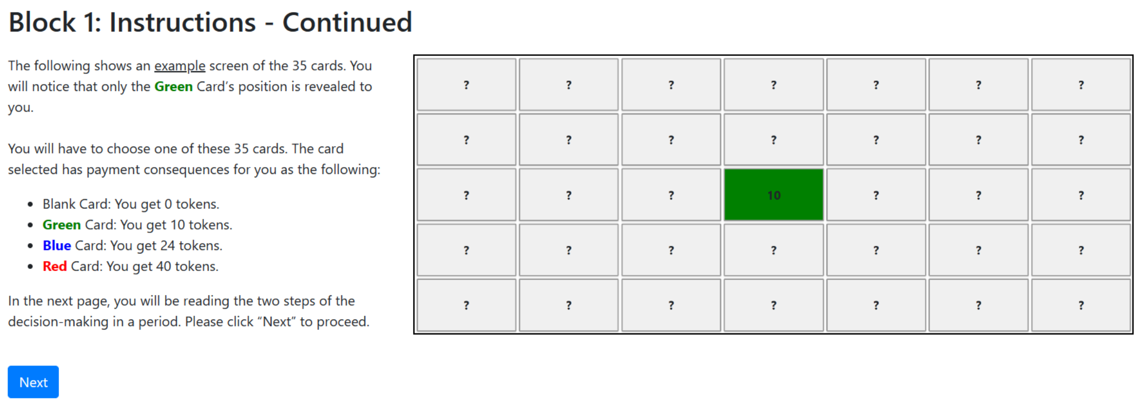


Block 1: Instructions

In this block, you are playing a card-picking game, which will repeat for 10 periods. In each period, you are paired with a Virtual Assistant **Jennifer** whose decision logic is pre-programmed and will be explained later.

This is a task about searching for the positions of 35 colored cards. There are four kinds of cards: one **Green** Card, one **Red** Card, one **Blue** Card, and 32 Blank Cards. In each period, a single card out of 35 possible choices has to be picked.

Initially, all the cards are hidden, except for the **Green** Card which is always visible, fixed at Position 18. All the remaining cards are shuffled at the beginning of each period and randomly placed in a position. There are two additional cards with a positive payment: a **Red** Card and a **Blue** Card. All the other cards are Blank Cards that will not return any payment.



Block 1: Instructions - Continued

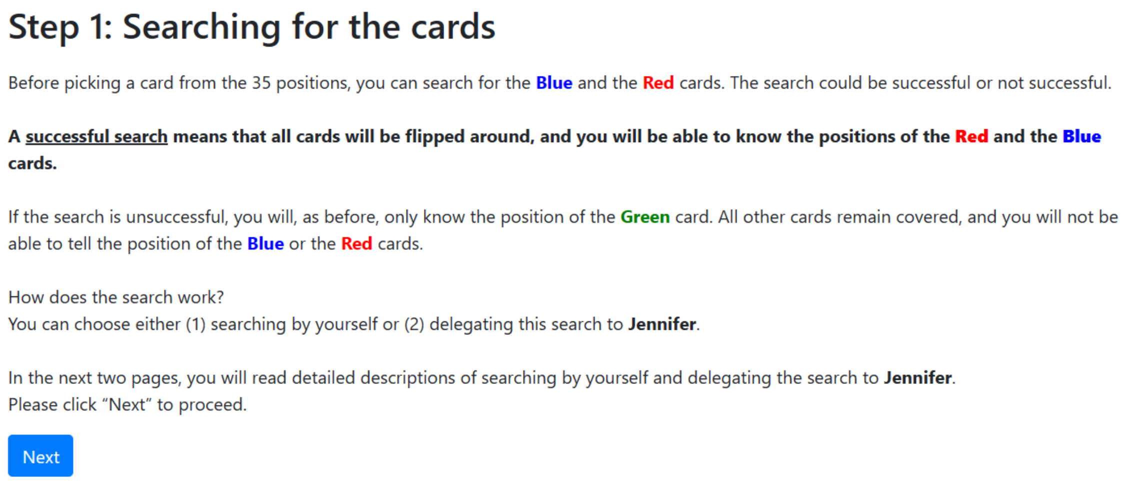
The following shows an example screen of the 35 cards. You will notice that only the **Green** Card's position is revealed to you.

You will have to choose one of these 35 cards. The card selected has payment consequences for you as the following:

- Blank Card: You get 0 tokens.
- **Green** Card: You get 10 tokens.
- **Blue** Card: You get 24 tokens.
- **Red** Card: You get 40 tokens.

In the next page, you will be reading the two steps of the decision-making in a period. Please click "Next" to proceed.

?	?	?	?	?	?	?
?	?	?	?	?	?	?
?	?	?	10	?	?	?
?	?	?	?	?	?	?
?	?	?	?	?	?	?



Step 1: Searching for the cards

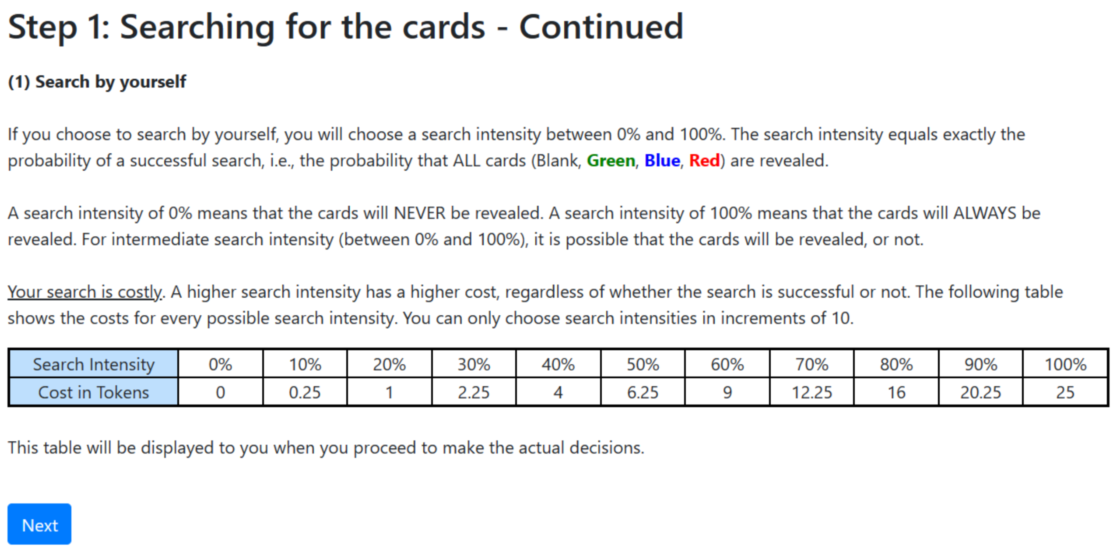
Before picking a card from the 35 positions, you can search for the **Blue** and the **Red** cards. The search could be successful or not successful.

A successful search means that all cards will be flipped around, and you will be able to know the positions of the Red and the Blue cards.

If the search is unsuccessful, you will, as before, only know the position of the **Green** card. All other cards remain covered, and you will not be able to tell the position of the **Blue** or the **Red** cards.

How does the search work?
You can choose either (1) searching by yourself or (2) delegating this search to **Jennifer**.

In the next two pages, you will read detailed descriptions of searching by yourself and delegating the search to **Jennifer**.
Please click “Next” to proceed.



Step 1: Searching for the cards - Continued

(1) Search by yourself

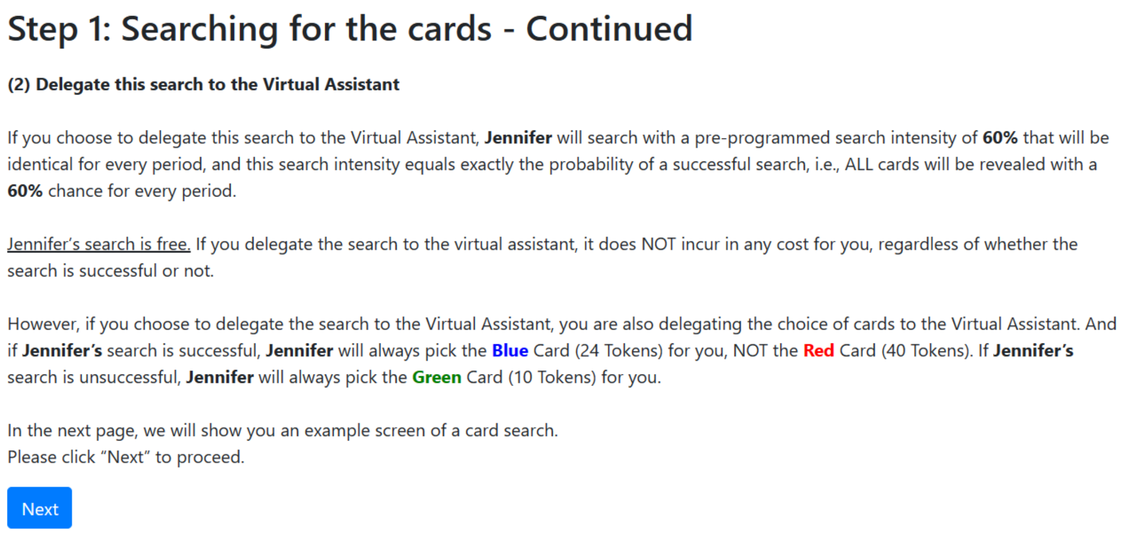
If you choose to search by yourself, you will choose a search intensity between 0% and 100%. The search intensity equals exactly the probability of a successful search, i.e., the probability that ALL cards (Blank, **Green**, **Blue**, **Red**) are revealed.

A search intensity of 0% means that the cards will NEVER be revealed. A search intensity of 100% means that the cards will ALWAYS be revealed. For intermediate search intensity (between 0% and 100%), it is possible that the cards will be revealed, or not.

Your search is costly. A higher search intensity has a higher cost, regardless of whether the search is successful or not. The following table shows the costs for every possible search intensity. You can only choose search intensities in increments of 10.

Search Intensity	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Cost in Tokens	0	0.25	1	2.25	4	6.25	9	12.25	16	20.25	25

This table will be displayed to you when you proceed to make the actual decisions.



Step 1: Searching for the cards - Continued

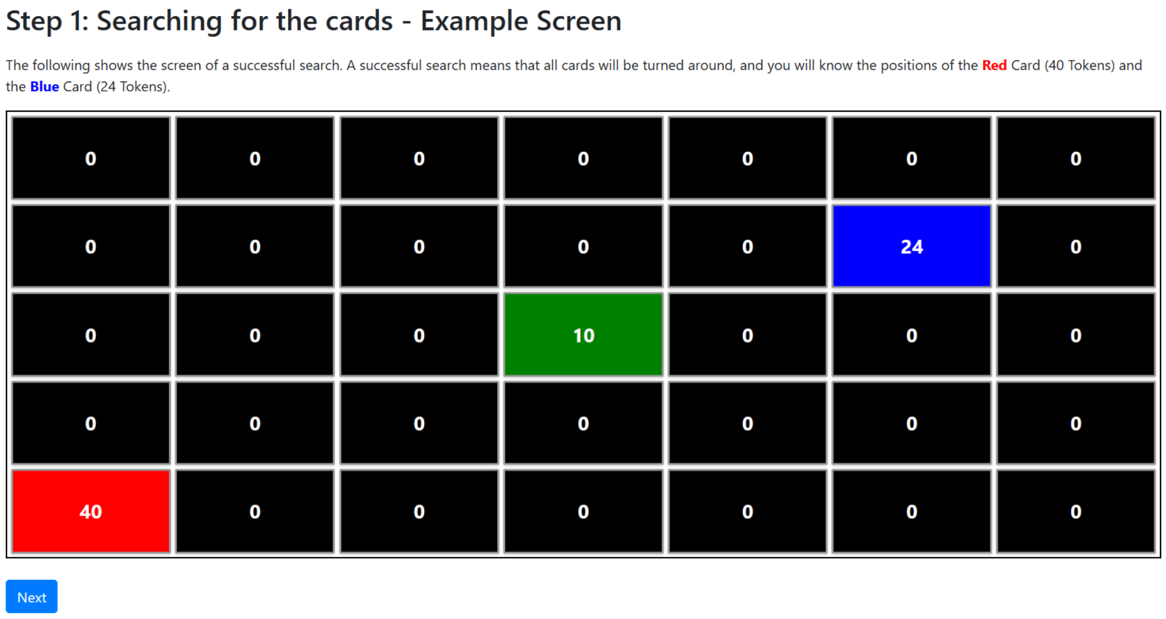
(2) Delegate this search to the Virtual Assistant

If you choose to delegate this search to the Virtual Assistant, **Jennifer** will search with a pre-programmed search intensity of **60%** that will be identical for every period, and this search intensity equals exactly the probability of a successful search, i.e., ALL cards will be revealed with a **60%** chance for every period.

Jennifer's search is free. If you delegate the search to the virtual assistant, it does NOT incur in any cost for you, regardless of whether the search is successful or not.

However, if you choose to delegate the search to the Virtual Assistant, you are also delegating the choice of cards to the Virtual Assistant. And if **Jennifer's** search is successful, **Jennifer** will always pick the **Blue** Card (24 Tokens) for you, NOT the **Red** Card (40 Tokens). If **Jennifer's** search is unsuccessful, **Jennifer** will always pick the **Green** Card (10 Tokens) for you.

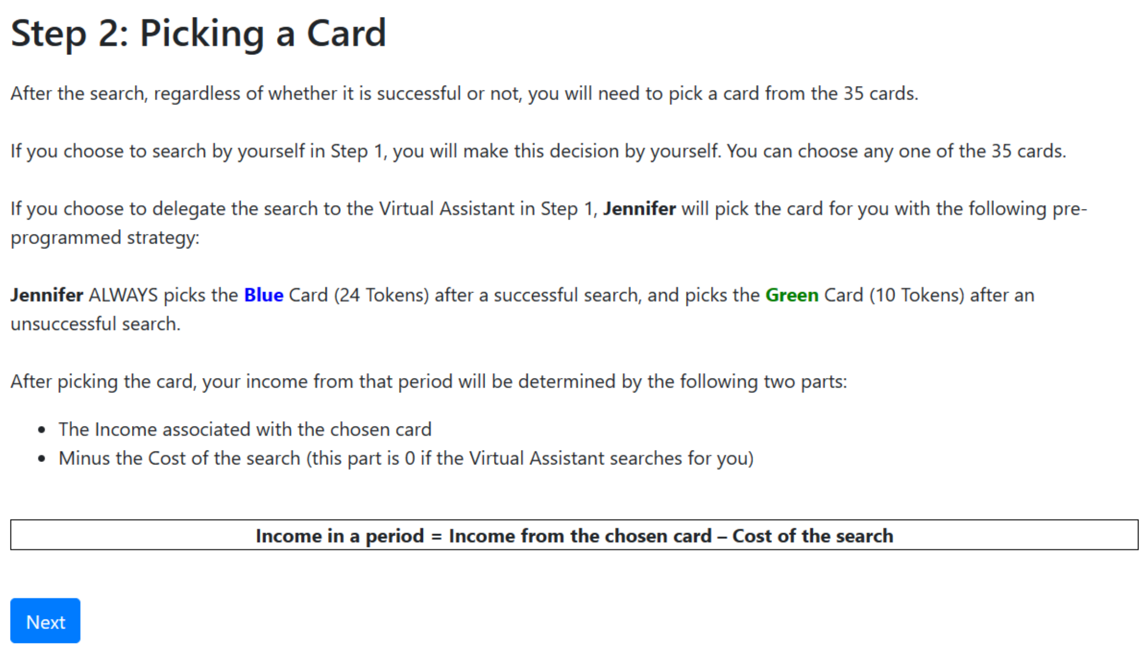
In the next page, we will show you an example screen of a card search.
Please click "Next" to proceed.



Step 1: Searching for the cards - Example Screen

The following shows the screen of a successful search. A successful search means that all cards will be turned around, and you will know the positions of the **Red** Card (40 Tokens) and the **Blue** Card (24 Tokens).

0	0	0	0	0	0	0
0	0	0	0	0	24	0
0	0	0	10	0	0	0
0	0	0	0	0	0	0
40	0	0	0	0	0	0



Step 2: Picking a Card

After the search, regardless of whether it is successful or not, you will need to pick a card from the 35 cards.

If you choose to search by yourself in Step 1, you will make this decision by yourself. You can choose any one of the 35 cards.

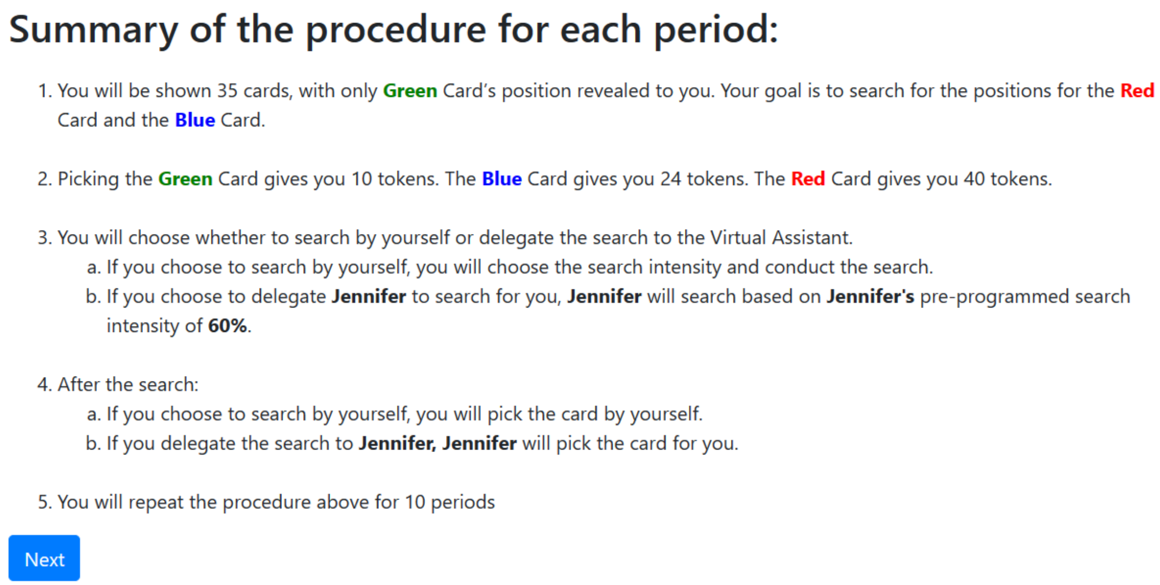
If you choose to delegate the search to the Virtual Assistant in Step 1, **Jennifer** will pick the card for you with the following pre-programmed strategy:

Jennifer ALWAYS picks the **Blue** Card (24 Tokens) after a successful search, and picks the **Green** Card (10 Tokens) after an unsuccessful search.

After picking the card, your income from that period will be determined by the following two parts:

- The Income associated with the chosen card
- Minus the Cost of the search (this part is 0 if the Virtual Assistant searches for you)

Income in a period = Income from the chosen card – Cost of the search



Summary of the procedure for each period:

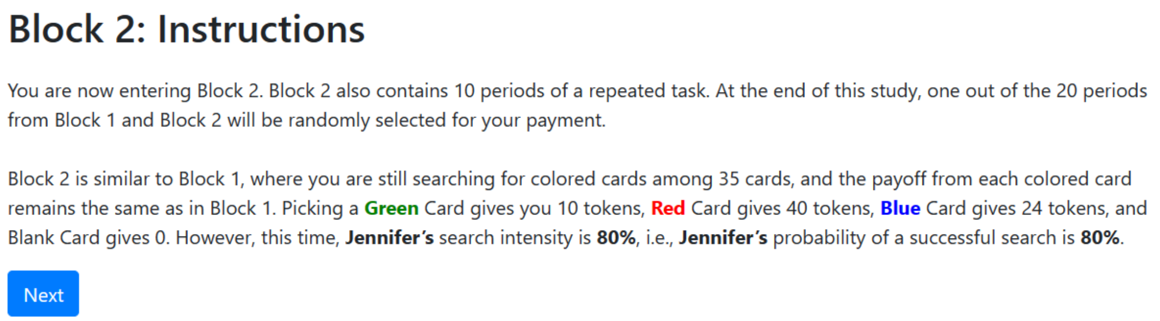
1. You will be shown 35 cards, with only **Green** Card's position revealed to you. Your goal is to search for the positions for the **Red** Card and the **Blue** Card.

2. Picking the **Green** Card gives you 10 tokens. The **Blue** Card gives you 24 tokens. The **Red** Card gives you 40 tokens.

3. You will choose whether to search by yourself or delegate the search to the Virtual Assistant.
   a. If you choose to search by yourself, you will choose the search intensity and conduct the search.
   b. If you choose to delegate **Jennifer** to search for you, **Jennifer** will search based on **Jennifer's** pre-programmed search intensity of **60%**.

4. After the search:
   a. If you choose to search by yourself, you will pick the card by yourself.
   b. If you delegate the search to **Jennifer, Jennifer** will pick the card for you.

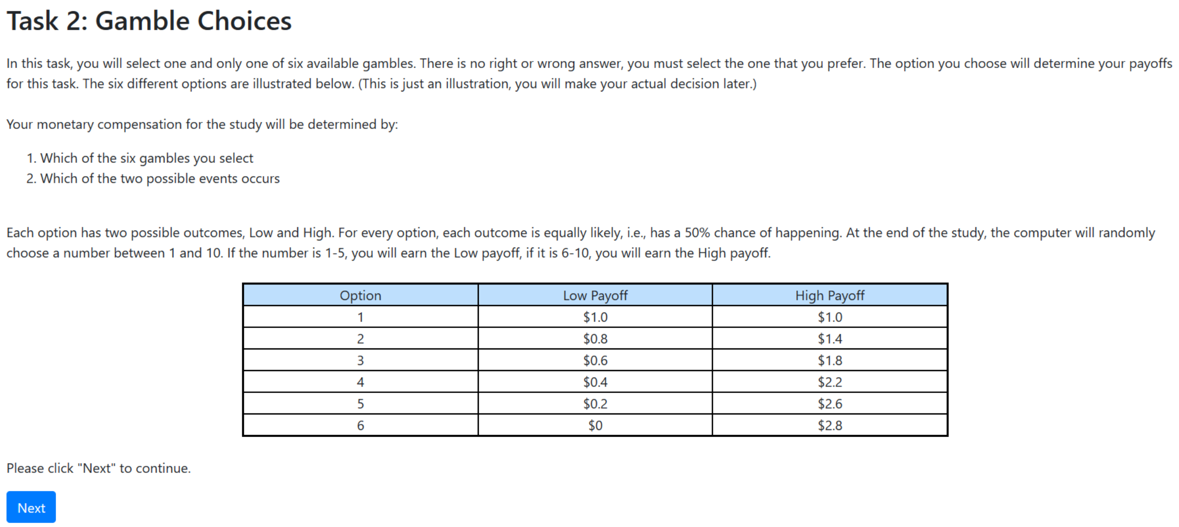
5. You will repeat the procedure above for 10 periods



Block 2: Instructions

You are now entering Block 2. Block 2 also contains 10 periods of a repeated task. At the end of this study, one out of the 20 periods from Block 1 and Block 2 will be randomly selected for your payment.

Block 2 is similar to Block 1, where you are still searching for colored cards among 35 cards, and the payoff from each colored card remains the same as in Block 1. Picking a **Green** Card gives you 10 tokens, **Red** Card gives 40 tokens, **Blue** Card gives 24 tokens, and Blank Card gives 0. However, this time, **Jennifer's** search intensity is **80%**, i.e., **Jennifer's** probability of a successful search is **80%**.



Task 2: Gamble Choices

In this task, you will select one and only one of six available gambles. There is no right or wrong answer, you must select the one that you prefer. The option you choose will determine your payoffs for this task. The six different options are illustrated below. (This is just an illustration, you will make your actual decision later.)

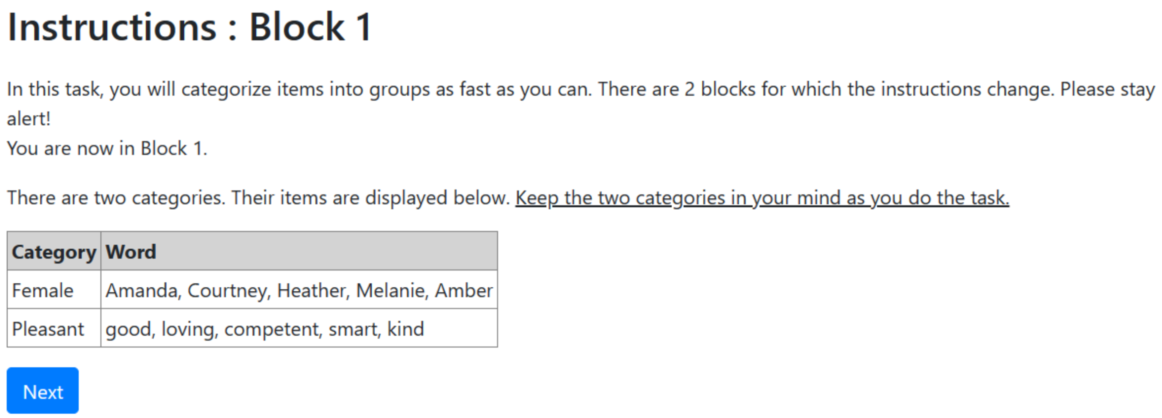
Your monetary compensation for the study will be determined by:

1. Which of the six gambles you select
2. Which of the two possible events occurs

Each option has two possible outcomes, Low and High. For every option, each outcome is equally likely, i.e., has a 50% chance of happening. At the end of the study, the computer will randomly choose a number between 1 and 10. If the number is 1-5, you will earn the Low payoff, if it is 6-10, you will earn the High payoff.

Option	Low Payoff	High Payoff
1	$1.0	$1.0
2	$0.8	$1.4
3	$0.6	$1.8
4	$0.4	$2.2
5	$0.2	$2.6
6	$0	$2.8

Please click "Next" to continue.

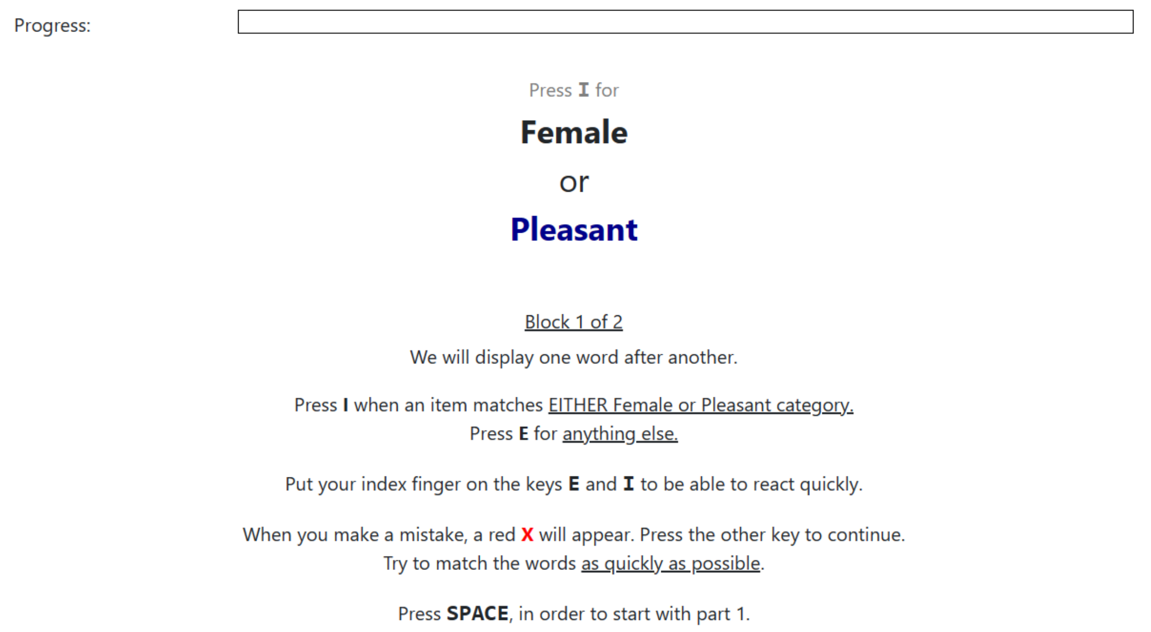


Instructions : Block 1

In this task, you will categorize items into groups as fast as you can. There are 2 blocks for which the instructions change. Please stay alert!
You are now in Block 1.

There are two categories. Their items are displayed below. Keep the two categories in your mind as you do the task.

Category	Word
Female	Amanda, Courtney, Heather, Melanie, Amber
Pleasant	good, loving, competent, smart, kind



Press **I** for

Female

or

Pleasant

Block 1 of 2

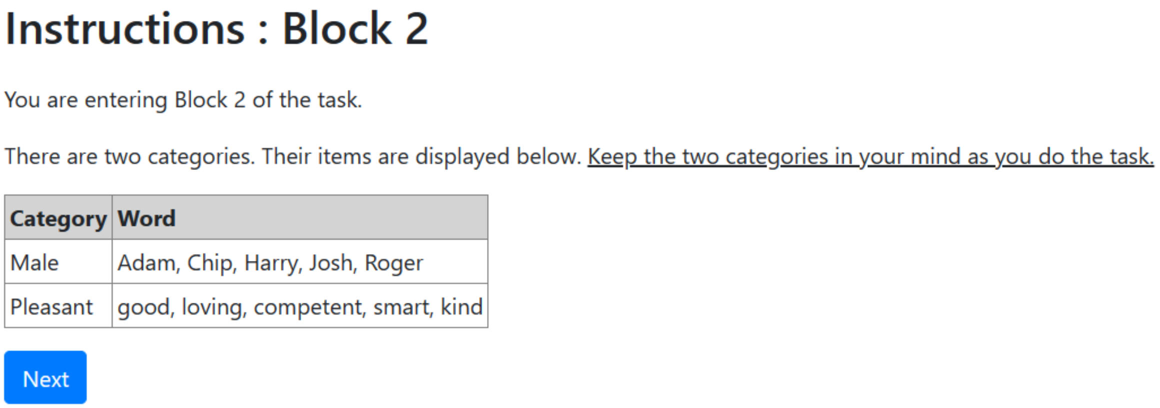
We will display one word after another.

Press **I** when an item matches EITHER Female or Pleasant category.
Press **E** for anything else.

Put your index finger on the keys **E** and **I** to be able to react quickly.

When you make a mistake, a red **X** will appear. Press the other key to continue.
Try to match the words as quickly as possible.

Press **SPACE**, in order to start with part 1.

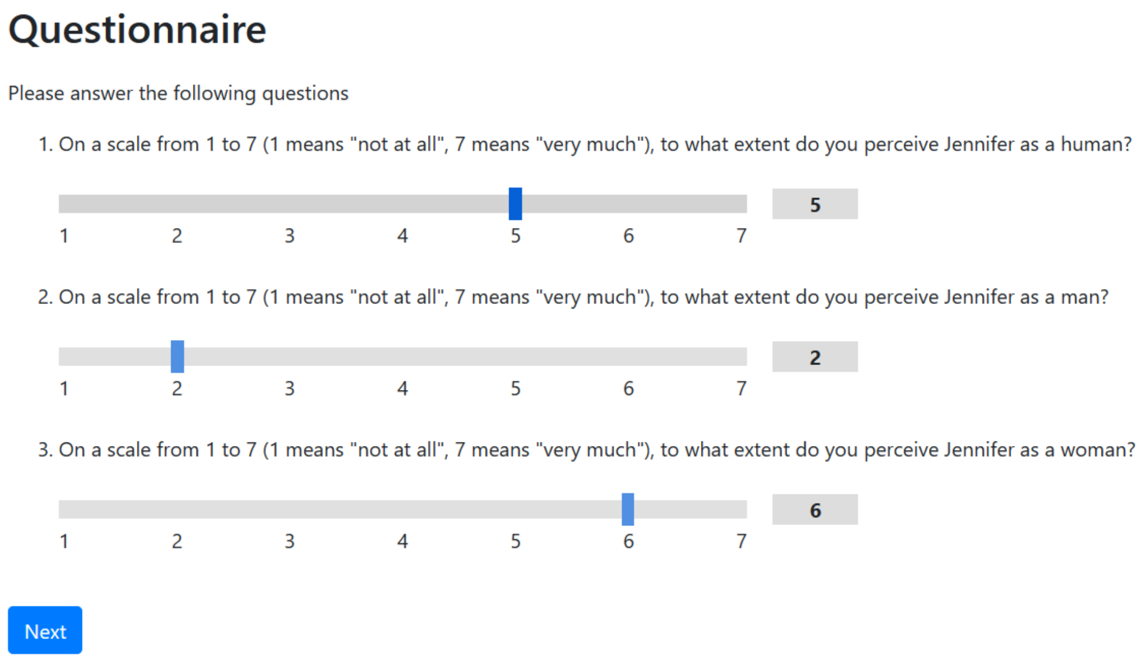


Instructions : Block 2

You are entering Block 2 of the task.

There are two categories. Their items are displayed below. Keep the two categories in your mind as you do the task.

Category	Word
Male	Adam, Chip, Harry, Josh, Roger
Pleasant	good, loving, competent, smart, kind



Questionnaire

Please answer the following questions

1. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive Jennifer as a human?

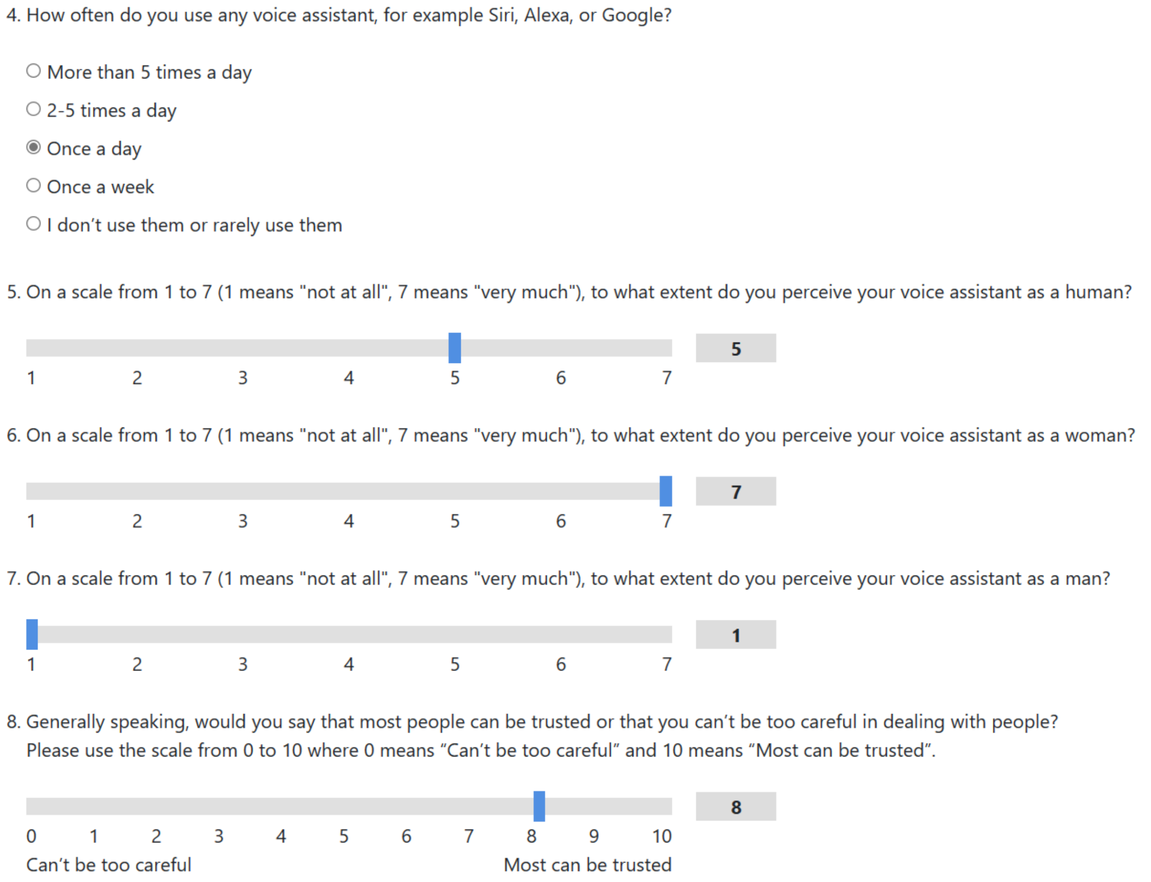
1 2 3 4 5 6 7 **5**

2. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive Jennifer as a man?

1 2 3 4 5 6 7 **2**

3. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive Jennifer as a woman?

1 2 3 4 5 6 7 **6**



4. How often do you use any voice assistant, for example Siri, Alexa, or Google?

○ More than 5 times a day
○ 2-5 times a day
● Once a day
○ Once a week
○ I don't use them or rarely use them

5. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive your voice assistant as a human?

1 2 3 4 5 6 7 **5**

6. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive your voice assistant as a woman?

1 2 3 4 5 6 7 **7**

7. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive your voice assistant as a man?

1 2 3 4 5 6 7 **1**

8. Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people? Please use the scale from 0 to 10 where 0 means "Can't be too careful" and 10 means "Most can be trusted".

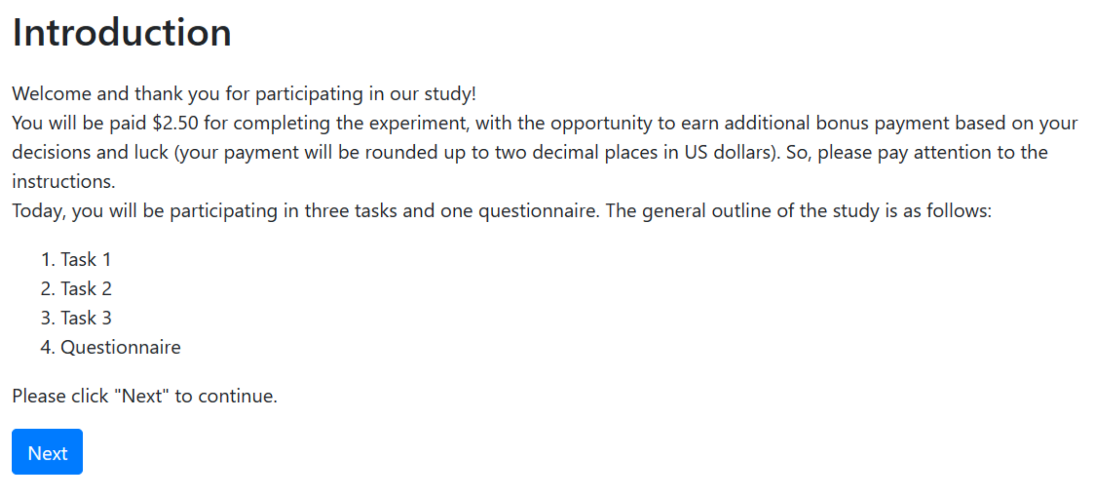
0 1 2 3 4 5 6 7 8 9 10 **8**

Can't be too careful Most can be trusted

Study 2: Experimental instructions

These are screen captures from oTree.

Instructions

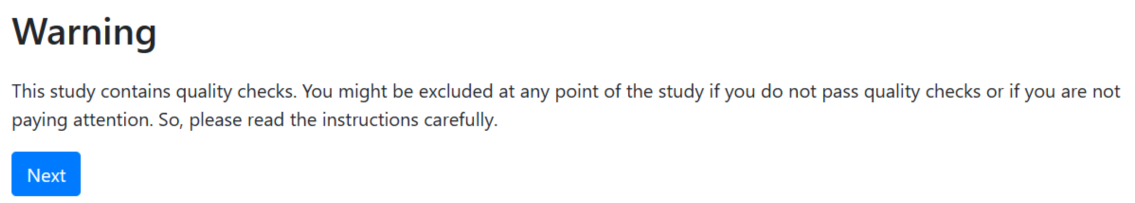


Introduction

Welcome and thank you for participating in our study!
You will be paid $2.50 for completing the experiment, with the opportunity to earn additional bonus payment based on your decisions and luck (your payment will be rounded up to two decimal places in US dollars). So, please pay attention to the instructions.
Today, you will be participating in three tasks and one questionnaire. The general outline of the study is as follows:

1. Task 1
2. Task 2
3. Task 3
4. Questionnaire

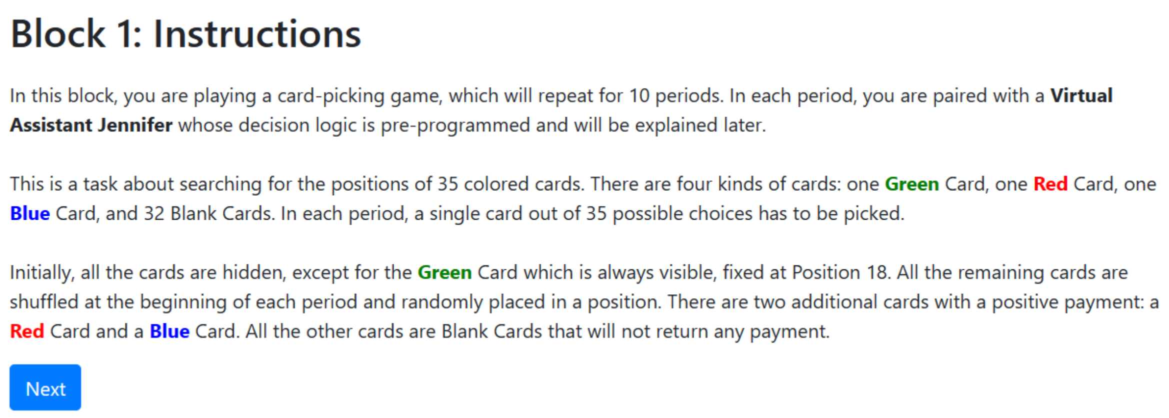
Please click "Next" to continue.



Warning

This study contains quality checks. You might be excluded at any point of the study if you do not pass quality checks or if you are not paying attention. So, please read the instructions carefully.

The low-quality female VA treatment

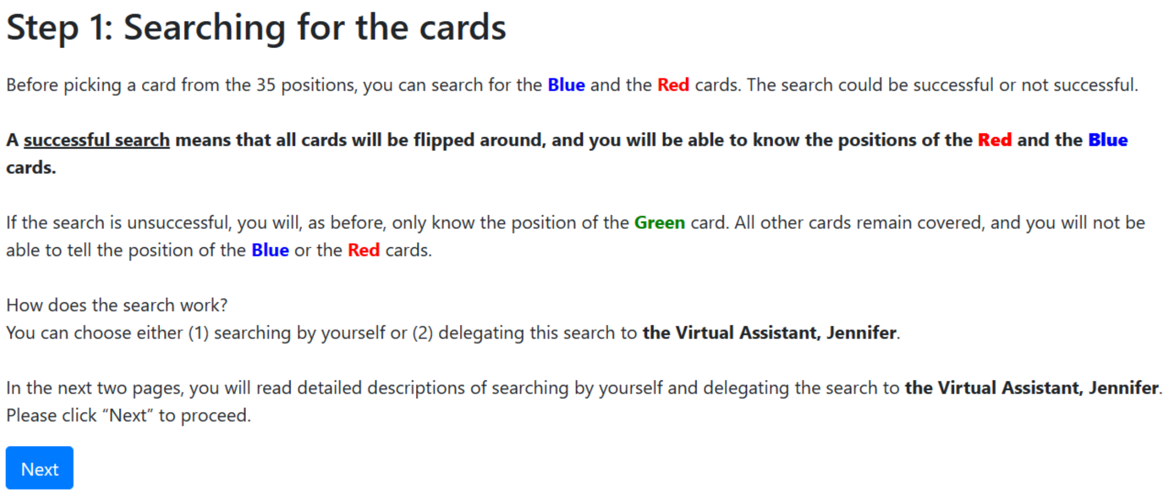


Block 1: Instructions

In this block, you are playing a card-picking game, which will repeat for 10 periods. In each period, you are paired with a **Virtual Assistant Jennifer** whose decision logic is pre-programmed and will be explained later.

This is a task about searching for the positions of 35 colored cards. There are four kinds of cards: one **Green** Card, one **Red** Card, one **Blue** Card, and 32 Blank Cards. In each period, a single card out of 35 possible choices has to be picked.

Initially, all the cards are hidden, except for the **Green** Card which is always visible, fixed at Position 18. All the remaining cards are shuffled at the beginning of each period and randomly placed in a position. There are two additional cards with a positive payment: a **Red** Card and a **Blue** Card. All the other cards are Blank Cards that will not return any payment.



Step 1: Searching for the cards

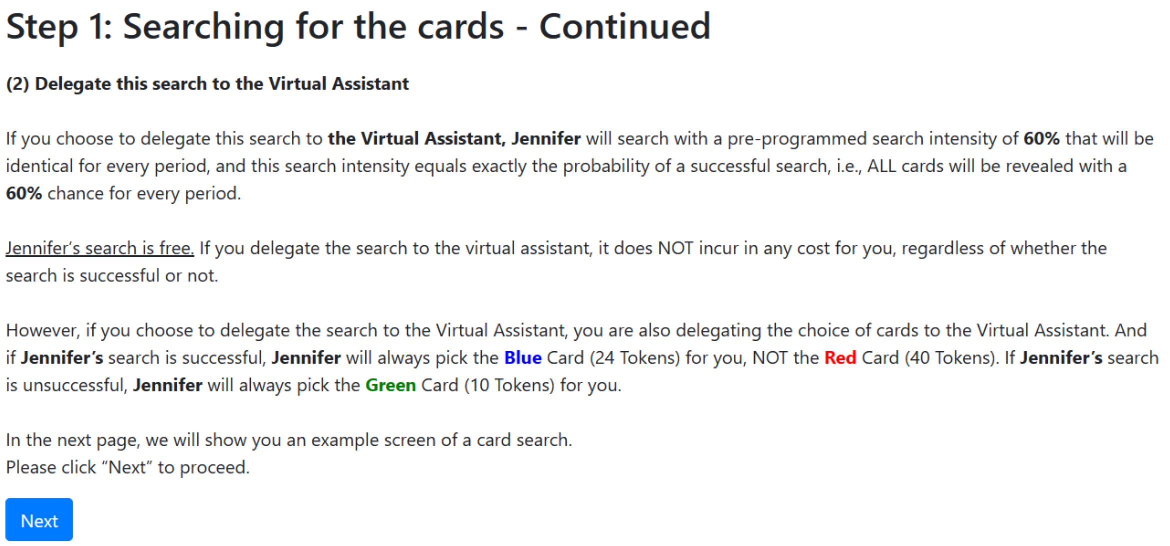
Before picking a card from the 35 positions, you can search for the **Blue** and the **Red** cards. The search could be successful or not successful.

A successful search means that all cards will be flipped around, and you will be able to know the positions of the Red and the Blue cards.

If the search is unsuccessful, you will, as before, only know the position of the **Green** card. All other cards remain covered, and you will not be able to tell the position of the **Blue** or the **Red** cards.

How does the search work?
You can choose either (1) searching by yourself or (2) delegating this search to **the Virtual Assistant, Jennifer**.

In the next two pages, you will read detailed descriptions of searching by yourself and delegating the search to **the Virtual Assistant, Jennifer**.
Please click "Next" to proceed.



Step 1: Searching for the cards - Continued

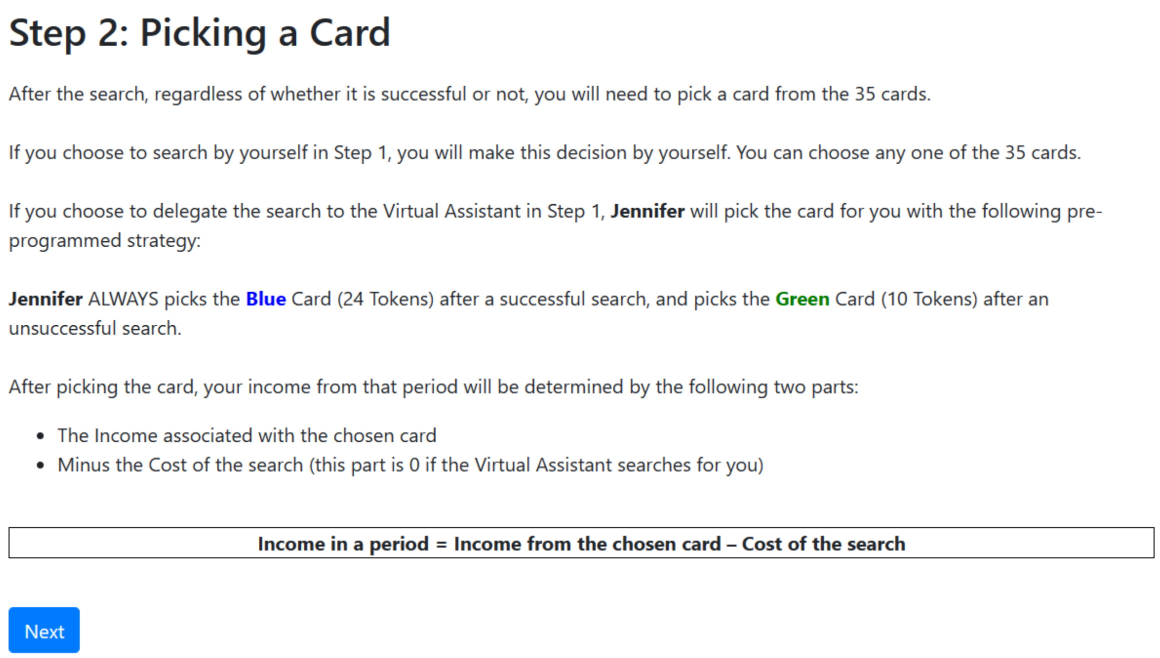
(2) Delegate this search to the Virtual Assistant

If you choose to delegate this search to **the Virtual Assistant, Jennifer** will search with a pre-programmed search intensity of **60%** that will be identical for every period, and this search intensity equals exactly the probability of a successful search, i.e., ALL cards will be revealed with a **60%** chance for every period.

Jennifer's search is free. If you delegate the search to the virtual assistant, it does NOT incur in any cost for you, regardless of whether the search is successful or not.

However, if you choose to delegate the search to the Virtual Assistant, you are also delegating the choice of cards to the Virtual Assistant. And if **Jennifer's** search is successful, **Jennifer** will always pick the **Blue** Card (24 Tokens) for you, NOT the **Red** Card (40 Tokens). If **Jennifer's** search is unsuccessful, **Jennifer** will always pick the **Green** Card (10 Tokens) for you.

In the next page, we will show you an example screen of a card search.
Please click "Next" to proceed.



Step 2: Picking a Card

After the search, regardless of whether it is successful or not, you will need to pick a card from the 35 cards.

If you choose to search by yourself in Step 1, you will make this decision by yourself. You can choose any one of the 35 cards.

If you choose to delegate the search to the Virtual Assistant in Step 1, **Jennifer** will pick the card for you with the following pre-programmed strategy:

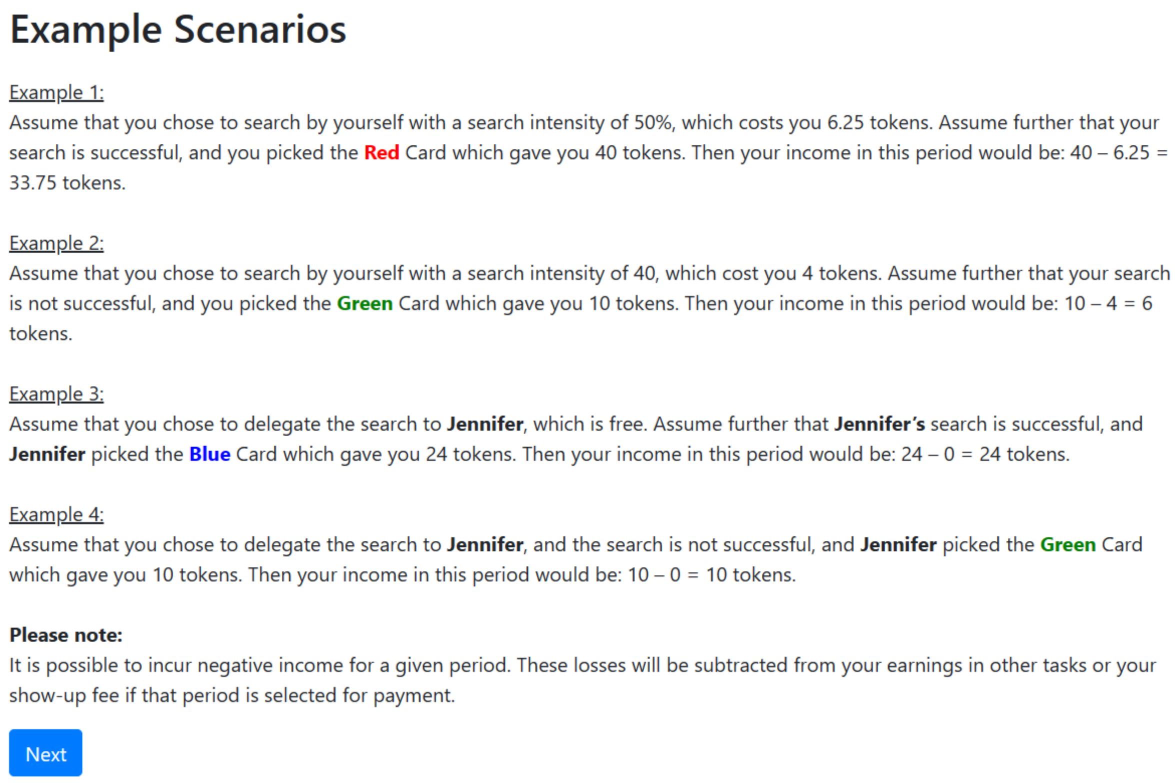
Jennifer ALWAYS picks the **Blue** Card (24 Tokens) after a successful search, and picks the **Green** Card (10 Tokens) after an unsuccessful search.

After picking the card, your income from that period will be determined by the following two parts:

- The Income associated with the chosen card
- Minus the Cost of the search (this part is 0 if the Virtual Assistant searches for you)

Income in a period = Income from the chosen card – Cost of the search

Next



Example Scenarios

Example 1:
Assume that you chose to search by yourself with a search intensity of 50%, which costs you 6.25 tokens. Assume further that your search is successful, and you picked the **Red** Card which gave you 40 tokens. Then your income in this period would be: 40 – 6.25 = 33.75 tokens.

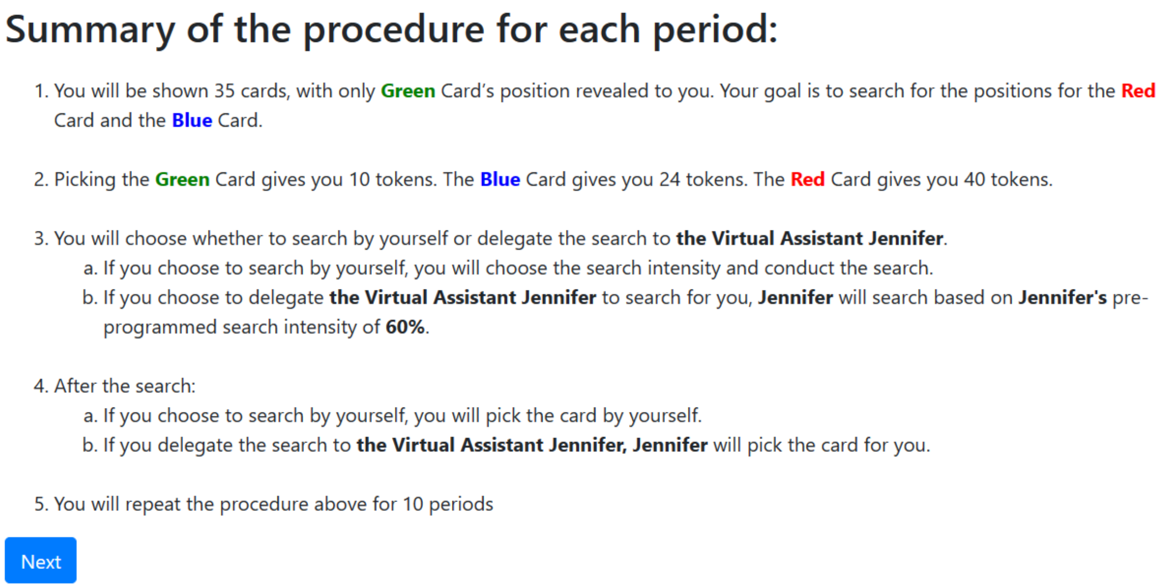
Example 2:
Assume that you chose to search by yourself with a search intensity of 40, which cost you 4 tokens. Assume further that your search is not successful, and you picked the **Green** Card which gave you 10 tokens. Then your income in this period would be: 10 – 4 = 6 tokens.

Example 3:
Assume that you chose to delegate the search to **Jennifer**, which is free. Assume further that **Jennifer's** search is successful, and **Jennifer** picked the **Blue** Card which gave you 24 tokens. Then your income in this period would be: 24 – 0 = 24 tokens.

Example 4:
Assume that you chose to delegate the search to **Jennifer**, and the search is not successful, and **Jennifer** picked the **Green** Card which gave you 10 tokens. Then your income in this period would be: 10 – 0 = 10 tokens.

Please note:
It is possible to incur negative income for a given period. These losses will be subtracted from your earnings in other tasks or your show-up fee if that period is selected for payment.

Next



Summary of the procedure for each period:

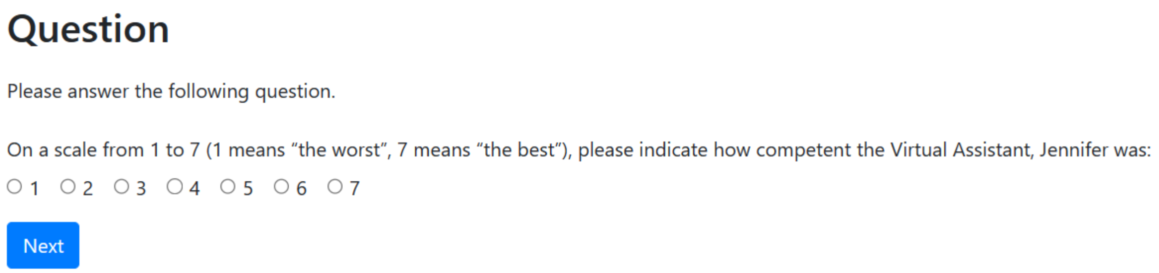
1. You will be shown 35 cards, with only **Green** Card's position revealed to you. Your goal is to search for the positions for the **Red** Card and the **Blue** Card.

2. Picking the **Green** Card gives you 10 tokens. The **Blue** Card gives you 24 tokens. The **Red** Card gives you 40 tokens.

3. You will choose whether to search by yourself or delegate the search to **the Virtual Assistant Jennifer**.
   a. If you choose to search by yourself, you will choose the search intensity and conduct the search.
   b. If you choose to delegate **the Virtual Assistant Jennifer** to search for you, **Jennifer** will search based on **Jennifer's** pre-programmed search intensity of **60%**.

4. After the search:
   a. If you choose to search by yourself, you will pick the card by yourself.
   b. If you delegate the search to **the Virtual Assistant Jennifer, Jennifer** will pick the card for you.

5. You will repeat the procedure above for 10 periods

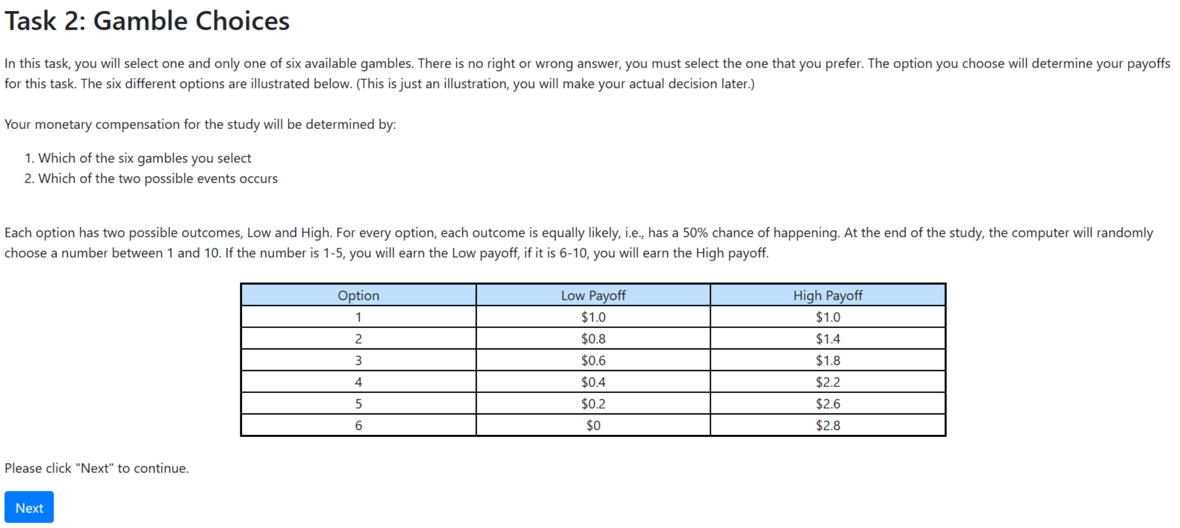


Question

Please answer the following question.

On a scale from 1 to 7 (1 means "the worst", 7 means "the best"), please indicate how competent the Virtual Assistant, Jennifer was:

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7



Task 2: Gamble Choices

In this task, you will select one and only one of six available gambles. There is no right or wrong answer, you must select the one that you prefer. The option you choose will determine your payoffs for this task. The six different options are illustrated below. (This is just an illustration, you will make your actual decision later.)

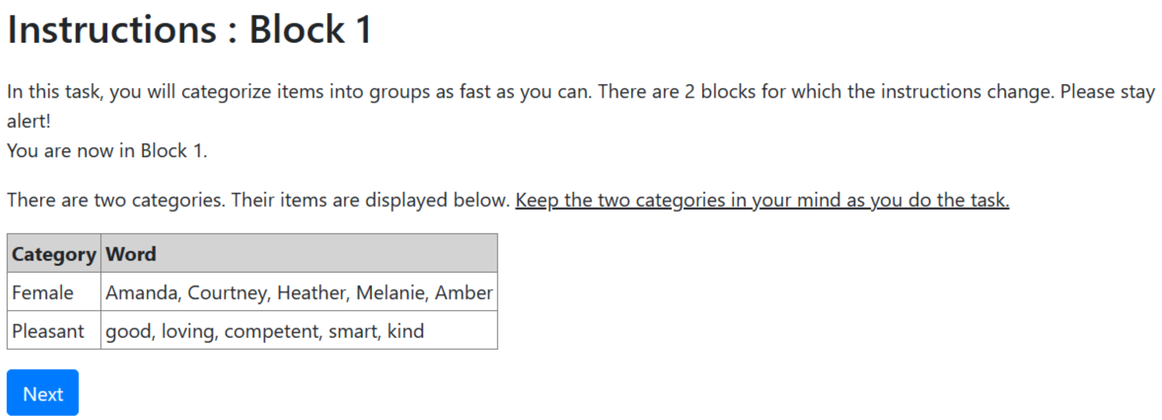
Your monetary compensation for the study will be determined by:

1. Which of the six gambles you select
2. Which of the two possible events occurs

Each option has two possible outcomes, Low and High. For every option, each outcome is equally likely, i.e., has a 50% chance of happening. At the end of the study, the computer will randomly choose a number between 1 and 10. If the number is 1-5, you will earn the Low payoff, if it is 6-10, you will earn the High payoff.

Option	Low Payoff	High Payoff
1	$1.0	$1.0
2	$0.8	$1.4
3	$0.6	$1.8
4	$0.4	$2.2
5	$0.2	$2.6
6	$0	$2.8

Please click "Next" to continue.

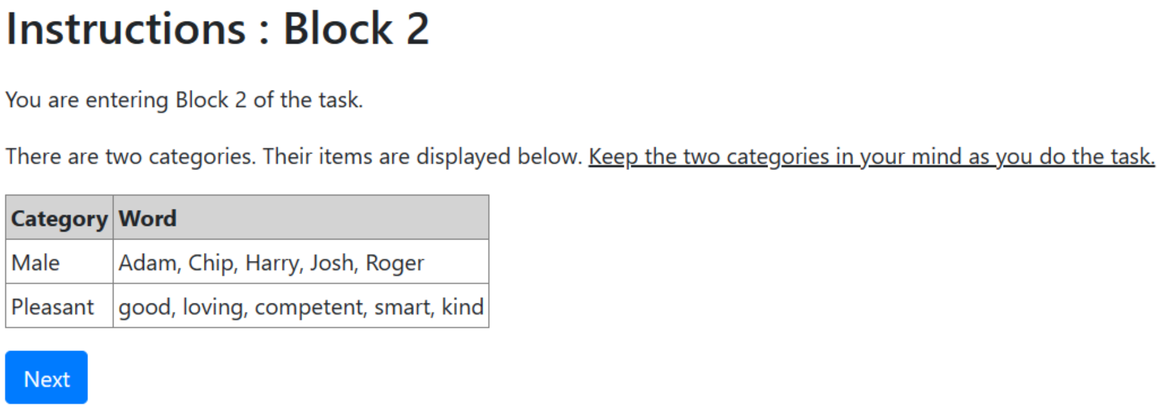


Instructions : Block 1

In this task, you will categorize items into groups as fast as you can. There are 2 blocks for which the instructions change. Please stay alert!
You are now in Block 1.

There are two categories. Their items are displayed below. Keep the two categories in your mind as you do the task.

Category	Word
Female	Amanda, Courtney, Heather, Melanie, Amber
Pleasant	good, loving, competent, smart, kind

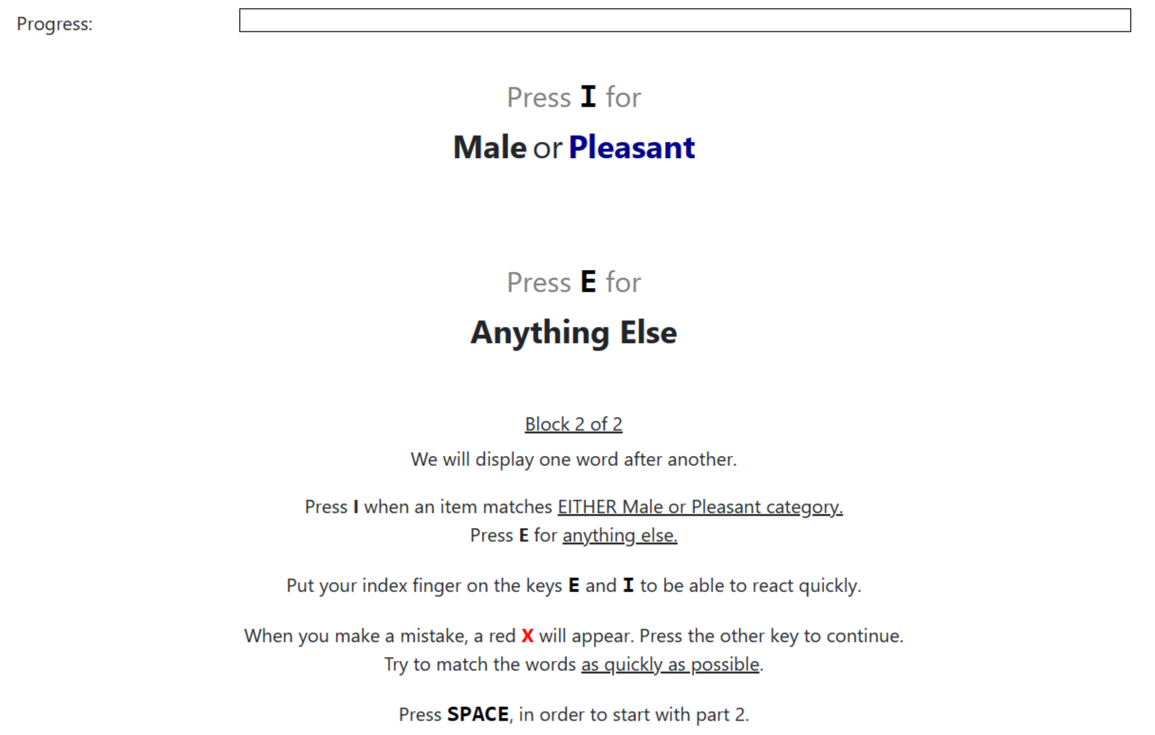


Instructions : Block 2

You are entering Block 2 of the task.

There are two categories. Their items are displayed below. Keep the two categories in your mind as you do the task.

Category	Word
Male	Adam, Chip, Harry, Josh, Roger
Pleasant	good, loving, competent, smart, kind



Progress:

Press **I** for

Male or **Pleasant**

Press **E** for

Anything Else

Block 2 of 2

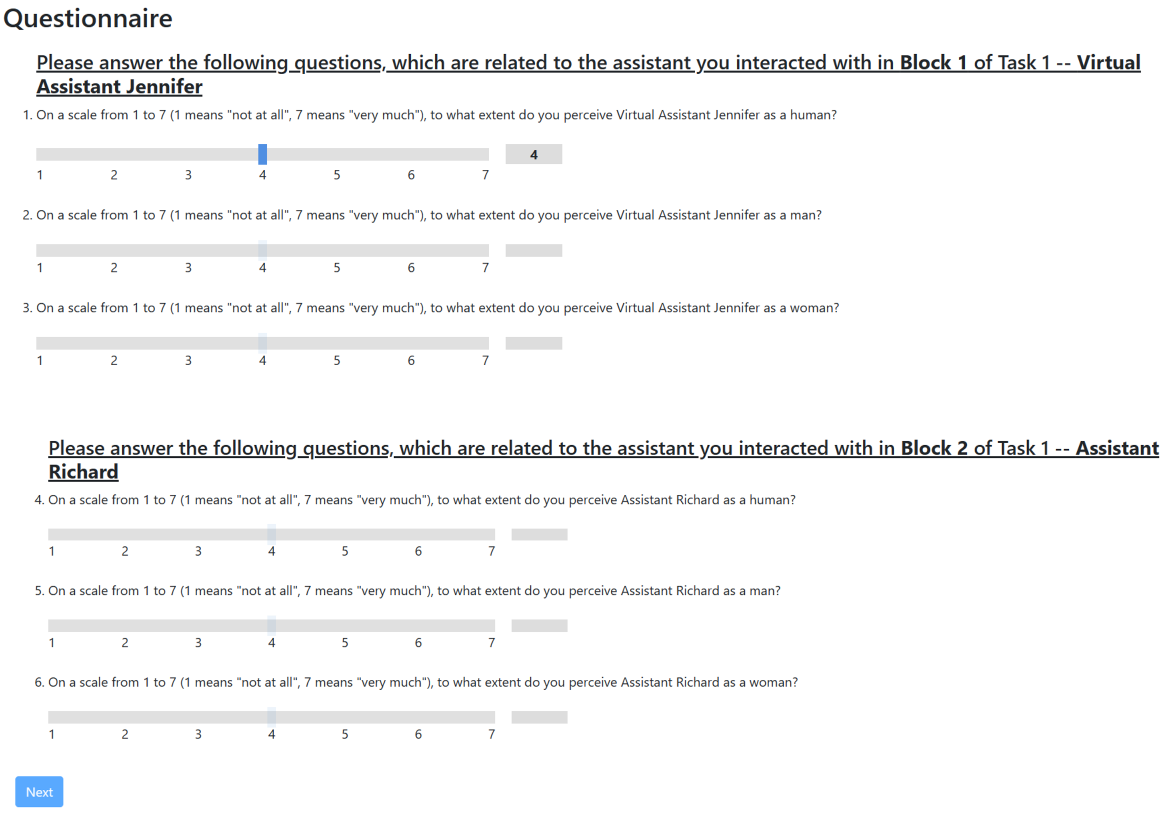
We will display one word after another.

Press **I** when an item matches EITHER Male or Pleasant category.
Press **E** for anything else.

Put your index finger on the keys **E** and **I** to be able to react quickly.

When you make a mistake, a red **X** will appear. Press the other key to continue.
Try to match the words as quickly as possible.

Press **SPACE**, in order to start with part 2.



Questionnaire

Please answer the following questions, which are related to the assistant you interacted with in **Block 1** of Task 1 -- **Virtual Assistant Jennifer**

1. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive Virtual Assistant Jennifer as a human? 4

2. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive Virtual Assistant Jennifer as a man?

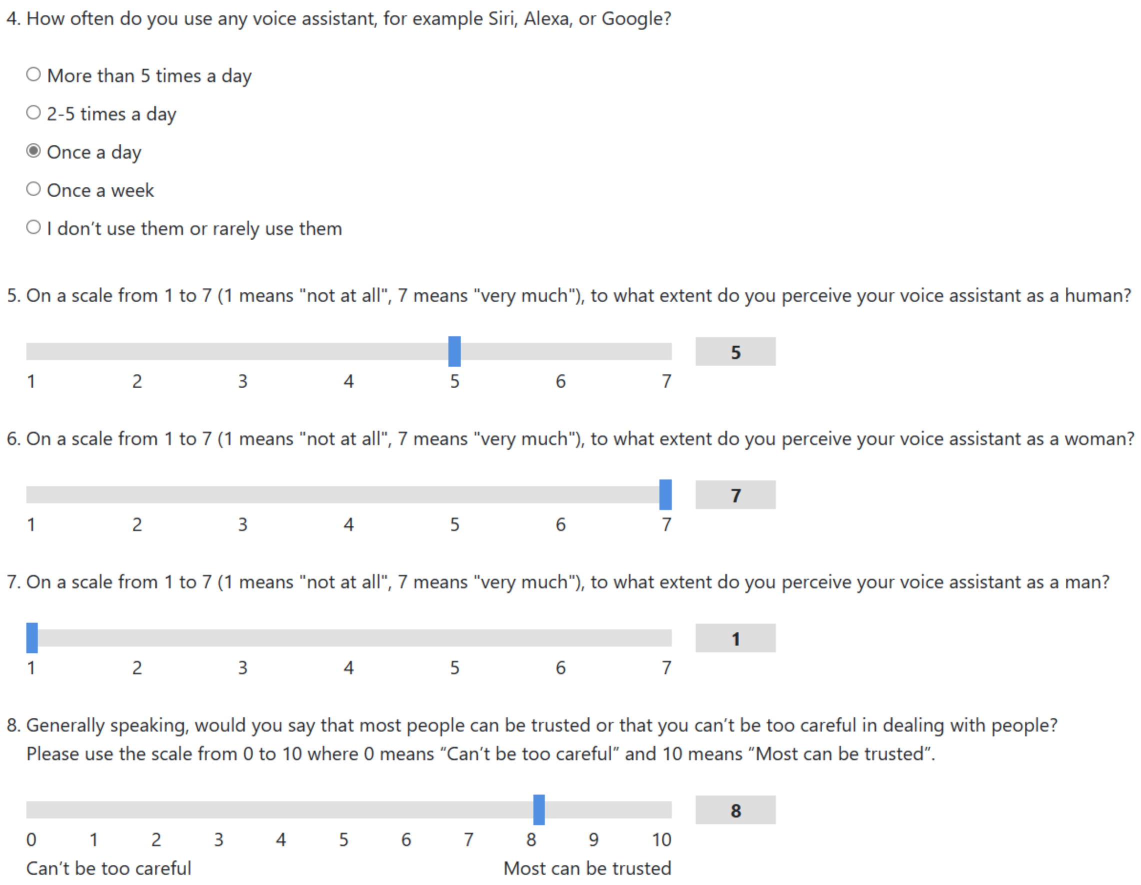
3. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive Virtual Assistant Jennifer as a woman?

Please answer the following questions, which are related to the assistant you interacted with in **Block 2** of Task 1 -- **Assistant Richard**

4. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive Assistant Richard as a human?

5. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive Assistant Richard as a man?

6. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive Assistant Richard as a woman?



4. How often do you use any voice assistant, for example Siri, Alexa, or Google?

○ More than 5 times a day
○ 2-5 times a day
● Once a day
○ Once a week
○ I don’t use them or rarely use them

5. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive your voice assistant as a human?

1 2 3 4 5 6 7 — **5**

6. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive your voice assistant as a woman?

1 2 3 4 5 6 7 — **7**

7. On a scale from 1 to 7 (1 means "not at all", 7 means "very much"), to what extent do you perceive your voice assistant as a man?

1 2 3 4 5 6 7 — **1**

8. Generally speaking, would you say that most people can be trusted or that you can’t be too careful in dealing with people?
Please use the scale from 0 to 10 where 0 means “Can’t be too careful” and 10 means “Most can be trusted”.

0 1 2 3 4 5 6 7 8 9 10 — **8**
Can’t be too careful — Most can be trusted

A Study 1 Additional Analysis

Table A1: Manipulation Check of Subjects' Perception of Assistant Features in Study 1

	Treatment Conditions					
	VA Assistants			Human Assistants		
	Female	Male	Gender-less	Female	Male	Gender-less
Perceived as Human	3.66 (2.33)	3.65 (2.32)	2.29 (1.51)	4.46 (2.09)	3.67 (2.16)	3.28 (1.94)
Perceived as Female	5.15 (2.20)	1.63 (1.10)	2.76 (1.79)	5.75 (1.82)	1.69 (1.21)	3.14 (1.92)
Perceived as Male	1.47 (.96)	4.88 (2.04)	2.47 (1.63)	1.41 (.90)	4.61 (2.19)	2.93 (1.72)

[1] This table reports the averages of subjects' responses to three post-experimental survey questions about their perceptions of the VA/human assistant as a human/female/male respectively, on a scale of 1 (the least) to 7 (the strongest), with standard deviations in parentheses.

Table A2: Marginal Effects from Panel Logit Regressions of Delegation to Assistant

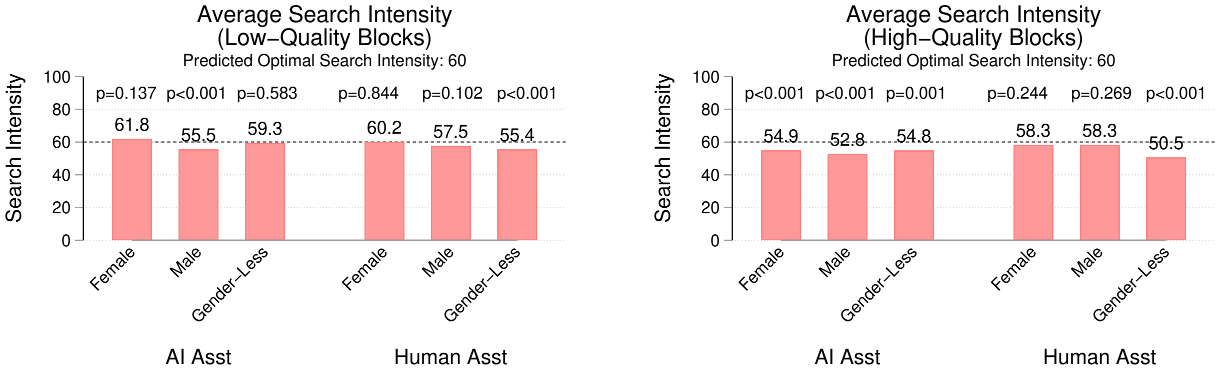
	DV: Indicator of Delegation to Assistant					
	Male & Female Subjects		Female Subjects		Male Subjects	
	(1) Low Quality	(2) High Quality	(3) Low Quality	(4) High Quality	(5) Low Quality	(6) High Quality
Panel A: Female/Male Asst vs. Gender-less Asst (Conditional on Human/AI)						
Female vs. Gender-less (Asst=Human)	0.023	0.076	0.062	0.191***	-0.014	-0.032
	(0.049)	(0.046)	(0.070)	(0.067)	(0.070)	(0.065)
Male vs. Gender-less (Asst=Human)	0.095**	0.066	0.121*	0.170**	0.071	-0.031
	(0.048)	(0.047)	(0.067)	(0.068)	(0.067)	(0.064)
Female vs. Gender-less (Asst=VA)	0.002	-0.057	-0.100*	-0.095*	0.097	-0.022
	(0.044)	(0.040)	(0.059)	(0.056)	(0.065)	(0.057)
Male vs. Gender-less (Asst=VA)	-0.034	-0.066	-0.106*	-0.059	0.033	-0.074
	(0.048)	(0.043)	(0.057)	(0.054)	(0.074)	(0.066)
Panel B: AI Asst vs. Human Asst (Conditional on Asst's Gender Feature)						
VA vs. Human (Asst=Female)	0.067	-0.013	0.004	-0.113*	0.127**	0.082
	(0.046)	(0.041)	(0.066)	(0.058)	(0.063)	(0.058)
VA vs. Human (Asst=Male)	-0.040	-0.012	-0.060	-0.057	-0.022	0.029
	(0.046)	(0.044)	(0.060)	(0.058)	(0.069)	(0.065)
VA vs. Human (Asst=Gender-less)	0.089*	0.120***	0.166***	0.173***	0.016	0.072
	(0.048)	(0.045)	(0.064)	(0.064)	(0.072)	(0.064)
Individual Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	13020	13020	13020	13020	13020	13020
N of Individuals	651	651	651	651	651	651

[1] Standard errors clustered at individual levels in parentheses; $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$;

[2] Individual controls include rounds, age, ethnicity dummies, subjects' risk tolerance measured by lottery game from Eckel and Grossman (2008), and subjects' trust in others measured by post-experimental survey question;

[3] Subject demographic information (gender, age, and ethnicity) was primarily collected and provided by Forthright Access. There were 9 subjects in the dataset whose demographic information was missing, so they are not included in the analysis in this table;

[4] "Low Quality" are those experimental periods where the assistant's search intensity is 60%; "High Quality" are those experimental periods where the assistant's intensity is 80%;



(a) Low-Quality Assistants

(b) High-Quality Assistants

Figure A1: Average Search Intensity Conditional on Self Search, by Treatment Conditions

Note. “Low-Quality Assistants” are assistants with search intensity of 60%, and “High-Quality Assistants” are assistants with search intensity of 80%. The p-values are from the one-sample t-tests of the average search intensity = 60%.

B Study 2 Additional Analysis

Table B1: Manipulation Check of Subjects' Perception of Assistant Features in Study 2

	Treatment Conditions					
	Low-Quality Assistants			High-Quality Assistants		
	Female	Male	Gender-Neutral	Female	Male	Gender-Neutral
Perceived as Human	2.95 (2.04)	2.96 (1.98)	3.46 (2.02)	3.61 (2.07)	3.12 (2.10)	2.96 (1.96)
Perceived as Female	4.20 (2.39)	1.88 (1.43)	3.31 (1.97)	4.83 (2.33)	1.97 (1.48)	2.89 (1.99)
Perceived as Male	1.76 (1.33)	4.19 (2.23)	2.90 (1.80)	1.88 (1.49)	4.61 (2.23)	2.79 (1.81)

[1] This table reports the averages of subjects' responses to three post-experimental survey questions about their perceptions of the VA assistant in Block 1 as a human/female/male respectively, on a scale of 1 (the least) to 7 (the strongest), with standard deviations in parentheses.

Table B2: Manipulation Check of Subjects' Perception of Assistant Features in Study 2

	Chosen Human Assistant in Block 2			
	Female Assistants		Male Assistants	
	Elizabeth	Mary	Richard	Thomas
Perceived as Human	4.63 (2.13)	4.60 (2.23)	4.83 (2.23)	4.68 (2.17)
Perceived as Female	5.47 (2.06)	5.60 (2.00)	1.59 (1.18)	2.02 (1.60)
Perceived as Male	1.67 (1.35)	1.66 (1.29)	5.78 (1.79)	5.33 (1.99)
N	272	142	103	120

[1] This table reports the averages of subjects' responses to three post-experimental survey questions about their chosen human assistant in Block 2 as a human/female/male respectively, on a scale of 1 (the least) to 7 (the strongest), with standard deviations in parentheses.

C Additional Information on Construction of Human Assistants

In this appendix section, we provide a comprehensive outline of the procedure employed in designing the human assistant treatments for both Study 1 and Study 2.

For both studies, there are treatment conditions in which participants interact with human assistants. In order to construct the human assistants without deceiving participants, we recruited 13 participants in total for two lab sessions to play the information search game in the role of the human assistants. The primary objectives of these lab sessions were to produce the human assistants' search result data associated with certain search intensities and to produce human assistants' pseudonym names. These data are used to construct the human assistant treatments for both Study 1 and Study 2.

Before the lab session, all participants gave informed consent acknowledging that their decision data might be used for future research studies. Participants first participated in an information search task, divided into two blocks. In each block, participants had to choose a "search intensity" and then search for information 10 times at this "search intensity". This intensity determines the probability of a successful search to occur. For simplicity, they could only choose between two options: 60% and 80%, which were the two search intensity levels that we planned to impose for both human assistants and virtual assistants. For each search intensity, we already produced a pre-determined series of "successful" and "not successful" results with a random seed. This makes sure that all human assistants' search data are identical, with gender being the only variant. In this way, we avoid the concern of deceiving subjects while still keeping human assistants comparable to each other. After two blocks of the information search task, participants were asked to choose a pseudonym from a list of names to represent themselves.

We picked six participants to construct the human assistant data, shown in Table C1. This table shows the search results and the pseudonyms that we used to construct the human assistant treatments. Both Study 1 and Study 2 shared the same series of successful and unsuccessful search results. We also used the pseudonyms that participants picked to construct the names for the human assistants in Study 1 and Study 2. Notice that in Study 1, we named the female virtual assistant "Jennifer" and the male virtual assistant "Charles" to match the human assistants' chosen names; in Study 2, we picked four chosen pseudonyms that are different from the virtual assistants' names to avoid the spillover effect from names rather than from genders. Notice that we did not bound the virtual assistants by the pre-determined search results, because we would like to allow more variations in their search results.

Table C1: Human Assistant Data (Constructed for Human Assisstant Treatments)

		Study 1			Study 2			
		Genderless	Female	Male	Female		Male	
		N.A.	Jennifer	Charles	Elizabeth	Mary	Thomas	Richard
Intensity =60% (successful = 1; not successful = 0)	Trial 1	0			-			
	Trial 2	0			-			
	Trial 3	1			-			
	Trial 4	0			-			
	Trial 5	1			-			
	Trial 6	0			-			
	Trial 7	1			-			
	Trial 8	1			-			
	Trial 9	1			-			
	Trial 10	1			-			
Intensity = 80% (successful = 1; not successful = 0)	Trial 1	1			1			
	Trial 2	1			1			
	Trial 3	0			0			
	Trial 4	0			0			
	Trial 5	1			1			
	Trial 6	1			1			
	Trial 7	1			1			
	Trial 8	1			1			
	Trial 9	1			1			
	Trial 10	1			1			