I Can Do Better Than Your AI: Expertise and Explanations

James Schaffer
US Army Research Laboratory
Playa Vista, CA
james.a.schaffer20.civ@mail.mil

John O'Donovan
University of California Santa Barbara
Santa Barbara, CA
jod@cs.ucsb.edu

James Michaelis
US Army Research Laboratory
Adelphi, MD
james.r.michaelis2.civ@mail.mil

Adrienne Raglin
US Army Research Laboratory
Adelphi, MD
adrienne.j.raglin.civ@mail.mil

Tobias Höllerer
University of California Santa Barbara
Santa Barbara, CA
holl@cs.ucsb.edu

ABSTRACT

Intelligent assistants, such as navigation, recommender, and expert systems, are most helpful in situations where users lack domain knowledge. Despite this, recent research in cognitive psychology has revealed that lower-skilled individuals may maintain a sense of illusory superiority, which might suggest that users with the highest need for advice may be the least likely to defer judgment. Explanation interfaces – a method for persuading users to take a system's advice – are thought by many to be the solution for instilling trust, but do their effects hold for self-assured users? To address this knowledge gap, we conducted a quantitative study (N=529) wherein participants played a binary decision-making game with help from an intelligent assistant. Participants were profiled in terms of both actual (measured) expertise and reported familiarity with the task concept. The presence of explanations, level of automation, and number of errors made by the intelligent assistant were manipulated while observing changes in user acceptance of advice. An analysis of cognitive metrics lead to three findings for research in intelligent assistants: 1) higher reported familiarity with the task simultaneously predicted more reported trust but less adherence, 2) explanations only swayed people who reported very low task familiarity, and 3) showing explanations to people who reported more task familiarity led to automation bias.

IUI '19, March 17–20, 2019, Marina del Ray, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302308

CCS CONCEPTS

• **Information systems** → **Expert systems**; Personalization; • **Human-centered computing** → User models; HCI design and evaluation methods.

KEYWORDS

User Interfaces; Human-Computer Interaction; Intelligent Assistants; Information Systems; Decision Support Systems; Cognitive Modeling

ACM Reference Format:
James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better Than Your AI: Expertise and Explanations. In *Proceedings of 24th International Conference on Intelligent User Interfaces (IUI '19).* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3301275.3302308

1 INTRODUCTION

Many people express distrust or negative sentiment towards intelligent assistants such as recommender systems, GPS navigation aids, and general purpose assistants (e.g., Amazon's Alexa). Simultaneously, the amount of information available to an individual decision maker is exploding, creating an increased need for automation of the information filtering process. Distrust may be due to intelligent assistants being overly complex or opaque, which generates uncertainty about their accuracy or relevance of results. This creates an enormous challenge for designers - even if intelligent assistants can deliver the right information at the right time, it can still be a challenge to persuade people to trust, understand, and adhere to their recommendations.

While trust is thought to be a primary issue in system acceptance, a user's self-assessment of his or her own knowledge may also play a significant role. Personality traits related to self-reported ability have been studied recently in cognitive psychology [33, 41]. A well known result from this research is the "Dunning-Kruger" effect, where low-ability individuals maintain an overestimated belief in their own

ability. This effect states that a lack of self-awareness, which is correlated with lower cognitive ability, may lead some to inflate their own skills while ignoring the skills of others. Moreover, the broader study of "illusory superiority," which tells us that it takes a highly intelligent person to gauge the intelligence of others, may play a role in the adoption of intelligent assistants. This begs the question: how can information systems sway over-confident users that believe they are smarter than the machine?

Explanations are becoming the commonly accepted answer to making intelligent assistants more usable (e.g, [29]). Explanations inform users about the details of automated information filtering (such as in recommender systems) and help to quantify uncertainty. Research into explanations started with expert systems studies [1, 28], and more recently their trust and persuasion effects have been studied in recommender systems [36, 50]. Beyond this, explanations have been shown to fundamentally alter user beliefs about the system, including competence, benevolence, and integrity [56]. On the negative side of things, theories of human awareness [17] and multitasking [23] predict that users who spend time perceiving and understanding explanations from intelligent assistants may have lower awareness of other objects in their environment (or must necessarily take more time to complete a given task), which could negatively affect performance [9]. This might have implications for "automation bias," or the over-trusting of technology [13], which occurs when users become complacent with delegating to the system. Whether explanations alleviate or exacerbate automation bias is still an open research question. There are still more intricacies to explanations, for instance, recent research in intelligent user interfaces [4] has found that explanations interact with user characteristics [25] to affect trusting beliefs about a recommender system, implying that not all explanations are equal for all users. Moreover, since intelligent assistants are often presented as autonomous, cognitive psychology might tell us that explanations that sway a cooperative and trusting user [7, 35] might be completely ignored by an overconfident individual. This problem could be amplified when imperfections in the systems are perceived.

The above concerns illustrate the idea that the use of explanations might need to be tailored to the user's self-confidence and the potential error of the information system or intelligent assistant. This work investigates how a person's prior reported familiarity with a task vs. their actual measured expertise affects their adherence to recommendations. A study of (N=529) participants is conducted on a binary choice task (the n-player Iterated Prisoner's Dilemma [2], here introduced in the form of the Diner's Dilemma [24]) with assistance from an experimentally manipulated intelligent assistant, dubbed the "Dining Guru." With this constrained setup, we were able to make theoretically ideal recommendations to the participants, or precisely control the level of error. We present an analysis of participant data that explains the consequences of self-assessed familiarity (our measure of confidence) for intelligent assistants. Changes in adherence to recommendations, trust, and situation awareness are observed. Our analysis allows us to address the following research questions:

(1) How does a person's self-assessed task knowledge predict adherence to advice from intelligent assistants?
(2) Are rational explanations from an intelligent assistant effective on over-confident people?
(3) What is the relationship between over-confidence, automation bias, system error, and explanations?

2 BACKGROUND

This section reviews the work that serves as the foundation of the experiment design and interpretation of participant data.

Explanations and Intelligent Assistants

The majority of intelligent assistant explanation research has been conducted in recommender and expert systems [49], and older forms of explanations present multi-point arguments (e.g., Toulmin style argumentation [32]) about why a particular course of action should be chosen. These explanations could be considered "rational" in that they explain the mathematical or algorithmic properties of the systems they explain, which contrasts with more emotional explanations, such as those that are employed by virtual agents [31, 38].The differential impact of explanation on novices and experts has also been studied [1]: novices are much more likely to adhere to recommendations due to a lack of domain knowledge, while expert users require a strong "domain-oriented" argument before adhering to advice. In this context, "expert" refers to a true expert (indicated by past experience) rather than a self-reported expert. Most of these studies focus on decision making domains (financial analysis, auditing problems), where success would be determined objectively. The work presented here builds on the research by Arnold et al. [1] by quantifying objectively determined expertise, and assessing its relationship with *self-reported* familiarity and interaction effects with system explanations.

Errors by Intelligent Assistants

Intelligent assistants often vary in their degree of automation and effectiveness. The pros and cons of varying levels of automation have been studied in human-agent teaming [8]. Less prompting of the user for intervention may reduce cognitive load, but might also reduce awareness of system operations. Although system effectiveness continues to improve (e.g. [39]), it is not conclusive that improved algorithms

result in improved adherence to system recommendations, for instance, no effect of system error was found in Salem et al. [52]. In contrast, Yu et al. found that system errors do have a significant effect on trust [57] and Harman et al. found that users trust inaccurate recommendations more than they should [30]. These research studies also call the relationship between trust and adherence into question. In this study, we attempt to clarify this relationship through simultaneous measurement of trust and adherence while controlling for system error and automation.

Self-reported Ability

Consequences of self-reported ability have been recently discovered in studies of cognitive psychology [33, 41]. As mentioned, the "Dunning-Kruger" effect predicts that low-ability individuals maintain an overestimated belief in their own ability. This work also illustrates how quantitative metrics collected through questionnaires do not always measure their face value. For instance, the Dunning-Kruger effect shows us that asking a user how much he knows about a particular topic will only quantify the number of "unknown unknowns" relative to the user, rather than their actual ability in that area [48]. Deficits in knowledge are a double burden for these users, not only causing them to make mistakes but also preventing them from realizing they are making mistakes [16, 42].

The Dunning-Kruger effect is part of a larger group of cognitive effects sometimes referred to as illusory superiority. Other effects in this category create an additional concern for the success of intelligent assistants. For instance, it is known that estimating the intelligence of others is a difficult task (the "Downing effect") [14] that requires high intelligence. This explains the tendency of people to be very likely to rate themselves as "above average," even though not everyone can be so. We might expect that lower-intelligence users would fail to accurately gauge the intelligence of information systems, leading to disuse.

The research on self-reported ability leads us to hypothesize that overconfident individuals are less likely to interact with or adhere to intelligent assistants, due to the overestimation of their own ability and their inability to assess the accuracy of the system. A main challenge in this work is quantifying illusory superiority for abstract trust games. Herein, we solicited users to self-report their familiarity with the games such as the Diner's Dilemma, Iterated Prisoner's Dilemma, or the Public Goods game beforehand. Anyone with a passing knowledge of the rules of this game should have an obvious advantage over their peers, for example, a commonly known fact among those familiar with trust games is that "tit-for-tat", which rewards cooperation with cooperation while consistently but not vengefully penalizing defection, is one of the best strategies to employ in the Iterated Prisoner's Dilemma.

Awareness and Automation Bias

Human awareness was introduced and is a focus of many US Department of Defense agencies [19], wherein it is referred to as "situation awareness." Situation awareness is defined as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future." Endsley's theory models human decision-making in dynamic environments and has become well accepted. There are three states of situation awareness, which are: perception, comprehension, and projection, all of which contribute to an individual's ability to make good decisions. Perhaps the biggest threat to situation awareness is automation bias [13], which predicts that the more of the information gathering process is automated, the more users will "check out." This is sometimes called "overtrusting" [34] and can, in some instances, lead to catastrophic failure[1]. Work on the Dunning-Kruger effect and automation bias led us to hypothesize that overconfident users would be less susceptible to automation bias, since it is likely they would prefer to perform the task on their own.

Trust

Despite the research on automation bias, research in intelligent assistants still points towards trust as an important factor to indicate usability. In the research literature, the term "trust" is used with many different semantic meanings and the way it is measured has implications for what it can predict. For instance, previous research in recommender systems has shown that trust and system perceptions are highly correlated [10, 36, 37, 40]. More generally, research by Mcknight ties the concepts of trust in technology and trust in other people [46], showing that people can discriminate between the two. In this work, we use a set of question items adapted from [47] for trust in a specific technology (in this case, the Dining Guru). This way, we can quantitatively assess the relationship between reported trust and observed adherence – a relationship that is often assumed

Iterated Prisoner's Dilemma (IPD) and Diner's Dilemma

The Diner's Dilemma was chosen for the research platform due to its wide applicability, limited complexity, and research base. The Diner's Dilemma is an n-player, iterated version of the basic Prisoner's Dilemma. In the basic Prisoner's

[1]https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk

	Player Chooses: Hotdog	Lobster
2 co-diners cooperate	20.00	24.00
1 co-diner cooperates	12.00	17.14
Neither cooperates	8.57	13.33

Table 1: Diner's Dilemma choice payoff matrix.

Dilemma, two players decide to take an action (cooperate or defect) without communication beforehand, where defection leads to a higher outcome for an individual regardless of the other players actions, but with mutual cooperation leading to a higher outcome than mutual defection. The iterated form of this game, Iterated Prisoner's Dilemma, is useful for studying many real-world situations such as military arms races, cooperation in business settings, economics, animal behavior, and doping in sports. Although motives to cooperate in the Iterated Prisoner's Dilemma can be dependent on many other factors such as altruistic tendencies, emotion, or the five-factor model [11, 15, 51], the game still represents a fairly simple binary choice task: in each round there are two options, with each option leading to a different objectively-measured outcome based on the current state of the game. Multi-player versions of this game, such as the Diner's Dilemma, are more complex, which has made them suitable for studying the effects of increased information available to players through a user interface [27, 45]. In this experiment, the 3-player version was used, which makes it sufficiently easy to understand but still sufficiently complex as to warrant a computational aid. Moreover, this version of the game has been previously validated in terms of understandability and information requirements [53].

3 GAME DESIGN

In the Diner's Dilemma, several diners eat out at a restaurant, repeatedly over an unspecified number of days, with the agreement to split the bill equally each time. Each diner has the choice to order the inexpensive dish (hotdog) or the expensive dish (lobster). Diners receive a better dining experience (here, quantified as *dining points*) when everyone chooses the inexpensive dish compared to when everyone chooses the expensive dish. To be a valid dilemma, the quality-cost ratio of the two items available in a valid Diner's Dilemma game must meet a few conditions. First, if the player were dining alone, ordering hotdog should maximize dining points. Second, players must earn more points when they are the sole defector than when all players cooperate. Finally, the player should earn more points when the player and the two co-diners all defect than when the player is the only one to cooperate. This "game payoff matrix" means that in one round of the game, individual diners are better off choosing the expensive dish regardless of what the others choose to do. However, over repeated rounds, a diner's choice can affect the perceptions of other co-diners and cooperation may develop, which affects long term prosperity of the group. Hotdog/lobster cost and values for the game are shown in Figure 1, under each respective item, resulting in the payoff matrix that is shown in Table 1.

Participants played the Diner's Dilemma with two simulated co-diners. The co-diners were not visually manifested as to avoid any confounding emotional responses from participants (see [11]). The co-diners played variants of Tit-for-Tat (TFT), a proven strategy for success in the Diner's Dilemma wherein the co-diner makes the same choice that the participant did in the previous round. To make the game more comprehensible for participants, simulated co-diners reacted only to the human decision and not to each other. In order to increase the information requirements of the game, some noise was added to the TFT strategy in the form of increased propensity to betray (respond to a hotdog order with a lobster order) or forgive (respond to a lobster order with a hotdog order). Participants played three games with an undisclosed number of rounds (approximately 50 per game) and co-diner strategies switched between games. This means that the primary task for the user was to figure out what strategies the co-diners were employing and adjust accordingly. In the first game, co-diners betrayed often and the best strategy was to order lobster. In the second game, co-diners betrayed at a reduced rate and also forgave to some degree, which made hotdog the best choice. In the final game, co-diners were very forgiving and rarely ordered lobster even when betrayed, which again made lobster the best choice. The mean performance of participants in each game is shown in Table4.

4 USER INTERFACE DESIGN

Participants played the game through the interface shown in Figure 1. This interface contains four components: the last round panel (left), the control panel (center), the history panel (bottom), and the virtual agent, the Dining Guru (right). Across treatments, all panels remained the same except for the Dining Guru, which varied (see Figure 2).

Participants were provided with a basic interface containing all of the information that the Dining Guru used to generate its advice. The last round panel was shown on the left side of the interface and the control panel was shown in the middle. Together, these panels displayed the current dining points, the food quality and cost of each menu item, the current round, and the results from the previous round in terms of dining points. These panels allowed the participant to make a choice in each round. On the lower portion of the screen a history panel was provided. This panel contained

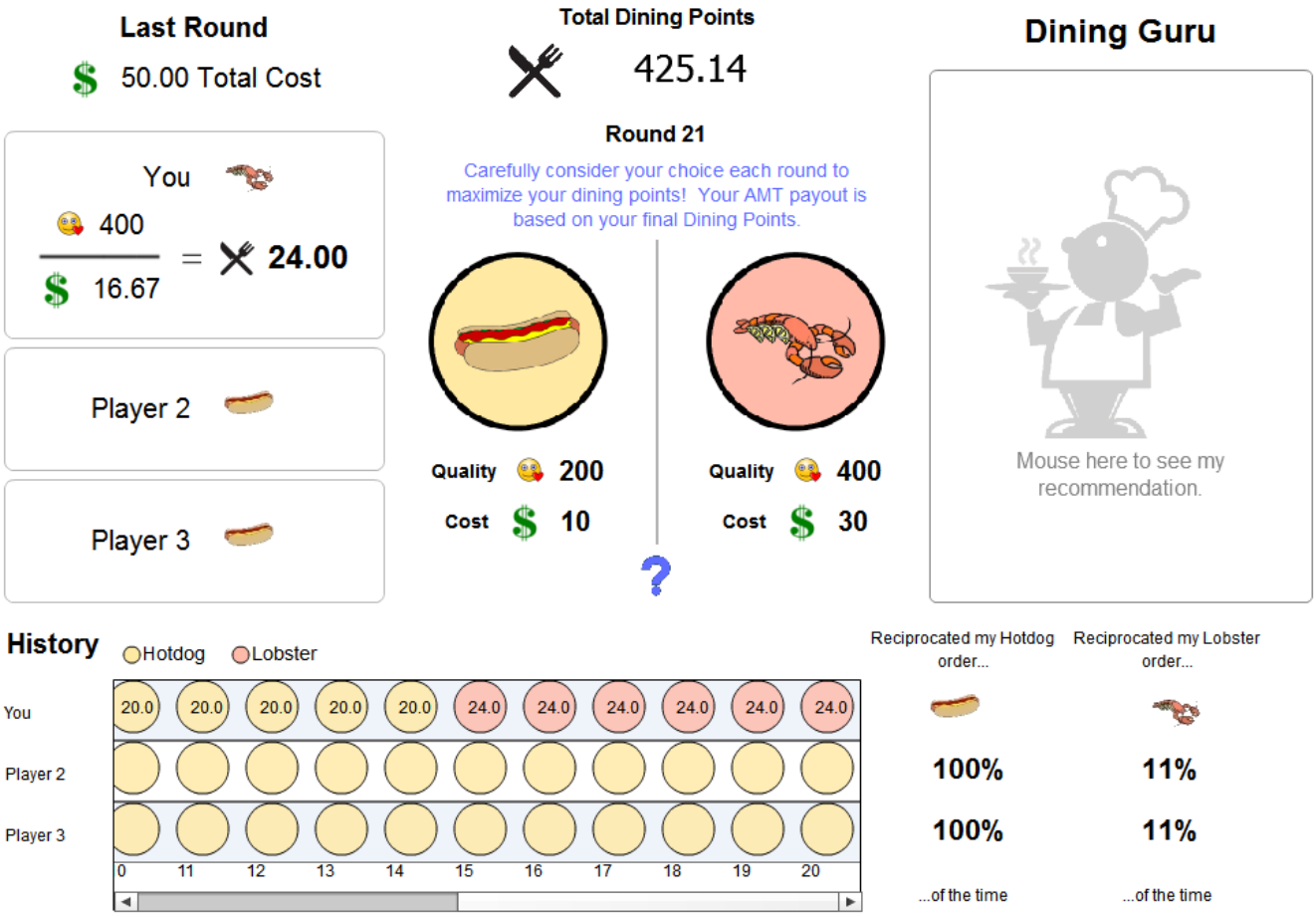

Figure 1: The user interface for the game.

information about who chose what in previous rounds and reciprocity rates.

Dining Guru

The Dining Guru was shown on the right side of the screen. In each round, the Dining Guru could be examined by a participant to receive a recommendation about which item (hotdog or lobster) would maximize their dining points. As with the simulated co-diners, the Dining Guru was not given any dynamic personality beyond being presented as an agent - a static drawing was used to communicate this. Users were required to mouse over the Dining Guru to invoke a recommendation, which made it possible to measure adherence. Recommendations were generated by calculating the expected value of ordering hotdog or lobster in the future, based on the maximum likelihood estimates of the rates of forgiveness and betrayal from co-diners. Due to the fixed strategy of the simulated co-diners, the Dining Guru made the "best possible" choice in each round, with most of the errors occurring in earlier rounds when information was incomplete. A "manual" version of the Dining Guru was given some treatments, which required participants to supply the Dining Guru with estimates of hotdog and lobster reciprocity rates ("Automation=0" cases in Figure 2) before receiving a recommendation. The rationale for this version is that by allowing the participant to become more engaged with the system, higher SA may result.

5 EXPERIMENT DESIGN

This section describes the experiment procedure, variables, and tested hypotheses.

Procedure

Reported familiarity was measured during the pre-study and trust was measured during the post-study (question items are given in Table 3). Before playing the game, participants were introduced to game concepts related to the user interface, reciprocity, and the Dining Guru by playing practice rounds.

Variable Name	μ	σ
Explanation (0 or 1)		
Automation (0 or 1)		
Error (0, 0.5, or 1)		
Self-reported Familiarity	0	1
Reported trust in the Dining Guru	0	1
Measured Expertise	8.08	1.64
Awareness of game elements	0.99	0.556
Adherence to system advice	0.33	0.25
Optimal Moves (as ratio, 0.0 to 1.0)	0.59	0.12

Table 2: A list of variables used in the analysis. Mean (μ) and standard deviation (σ) are given for dependent variables. Self-reported familiarity and reported trust were strictly latent factors, thus having a mean of 0 and standard deviation of 1. Awareness and measure expertise were parcelled (averaged) into latent factors before analysis.

Self-rep. Familiarity ($\alpha = 0.80, AVE = 0.59$)	R^2	*Est.*
I am familiar with abstract trust games.	0.52	1.27
I am familiar with the Diner's Dilemma.	0.47	1.14
I am familiar with the public goods game.	0.76	1.58
Trust ($\alpha = 0.9, AVE = 0.76$)	R^2	*Est.*
I trusted the Dining Guru.	0.79	1.36
I could rely on the Dining Guru.	0.74	1.29
I would advise a friend to take advice from the Dining Guru if they played the game.	0.75	1.37

Table 3: Factors determined by participant responses to subjective questions. R^2 reports the fit of the item to the factor. Est. is the estimated loading of the item to the factor. α is Cronbach's alpha (a measurement of reliability, see [54]). AVE is average variance extracted.

Several training questionnaires, which could be resubmitted as many times as needed, were used to help participants learn the game. The Dining Guru was introduced as an "AI adviser" and participants learned how to access it and what its intentions were. Participants were told that the Dining Guru was not guaranteed to make optimal decisions and that choosing to take its advice was their choice. Expertise was measured just after training by assessing retention of game and user interface concepts via an eleven item questionnaire. During the primary game phase, participants played three games of Diner's Dilemma against three configurations of simulated co-diners with varying behavior characteristics.

Independent Variables

Two levels of automation ($Automation = 0$, $Automation = 1$), two levels of explanation ($Explanation = 0$, $Explanation = 1$) and three levels of recommendation error ($Error = 0$, $Error = 0.5$, $Error = 1$) were manipulated between-subjects (see Figure 2 for a visual). All manipulations (three parameters, $3 * 2^2 = 12$ manipulations) were used as between-subjects treatments in this experiment.

The explanation for the Dining Guru was designed to accurately reflect the way that it was calculating recommendations. This was to align it with other forms of "rational" explanation [32], such as collaborative filtering explanations [49]. Since the Dining Guru calculates maximum-likelihood estimates of co-diner behavior and cross references this with the payoff matrix to produce recommendations, the explanation thus needed to contain estimates for the expected points per round of each choice. Additionally, in the manual version, in which the human player would provide reciprocation rate estimates to the dining guru, a text blurb appeared explaining the connection between co-diner reciprocity rates and the expected per-round average. The explanatory version relayed the expected per-round averages graphically, in a bar chart, right above the recommendation, so that participant attention would be drawn to the explanation. The explanation variable thus represents whether or not the system was a "black-box" ($Explanation = 0$) or a "white-box" ($Explanation = 1$) for each participant.

In the "manual" treatment ($Automation = 0$), the Dining Guru did not update until prompted by the user, who was required to provide estimates of co-diners' reciprocity rates. The estimates were provided by moving two sliders, which displayed and took no value until users first interacted with them. Users could freely experiment with the sliders, which means that they could be used to understand the relationship between the payoff matrix and co-diner reciprocity rates.

Three levels of error were manipulated: no-error, weak-error and full-error. In the no-error treatment ($Error = 0.0$), the Dining Guru produced recommendations that could be considered flawless, which, if followed, would result in optimal moves. The weak error ($Error = 0.5$) version would randomly adjust the reciprocity estimates up or down by up to 25%. For instance, if the true hotdog reciprocity rate was 65%, the Dining Guru would use a value anywhere between 40 and 90%. This could cause the Dining Guru to "flip-flop" between recommendations. Finally, the "full" error ($Error = 1.0$) condition adjusted reciprocity estimates by up to 50% in either direction. A practical consequence of this was that the Dining Guru would flip its recommendation almost every round. The error in the recommendations was reasonably hidden from participants and indeed was only noticeable when either explanation was present or the Dining Guru was not automated.

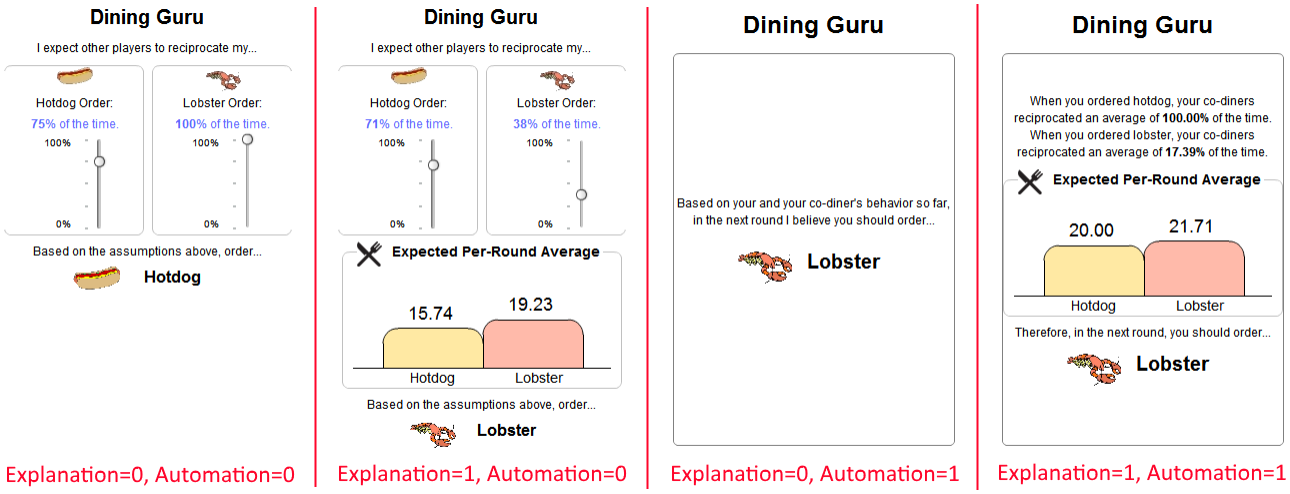

Figure 2: Variations of the Dining Guru, based on automation and explanation treatments.

Dependent Variables

Adherence was said to occur for each round where the user choice matched the last recommendation given by the Dining Guru. The measurement of true adherence is a difficult research challenge, since the reason for a participant's decision following a recommendation must be the recommendation itself, and not because the participant would have chosen that regardless. The mouse-over design helped to reduce some of the uncertainty about the reason for the participant's choice. The adherence measurement was scaled between 0 and 1.

Performance, measured as the ratio of moves that were considered optimal, was analyzed on a per-participant basis for the purpose of validating the Dining Guru's design. While learning effects were found in the data, they are not the focus of this paper and were not involved in our statistical tests (discussion of learning effects in the Diner's Dilemma game can be found in [53]).

Reported familiarity was chosen to quantify illusory superiority - this was done for three reasons. First, agreement questions such as "I am an expert Iterated Prisoner's Dilemma player" would not have measured the desired factor - outside of tournaments where researchers submit *algorithms* play the IPD[2], there is, to our knowledge, no competitive community of Diner's Dilemma or IPD players. Second, the "Dunning-Kruger" effect is demonstrated by measuring "unknown-unknowns," and likert-scale based questions about familiarity measures "I don't even know what this is" all the way to "thorough mastery," as the common dictionary definition suggests. Third, Diner's Dilemma, similar to chess, is a game about information and decision-making. For instance, you would expect someone who is intimately familiar with the rules of chess to perform better than someone who does not know how the pieces move, but this is not true for people who may report being very familiar with baseball.

A factor analysis of trust and self-reported familiarity is given in Table 3. All of these items were taken on a Likert scale (1-7), which rated agreement. Furthermore, we demonstrate the external validity of these factors by examining correlations with both measured expertise and performance – if the self-reported familiarity metric is valid, negative correlations should be found for both. All factors achieved discriminant validity using the Campbell & Fiske test [6].

Situation awareness and (true) expertise were measured using testing methodologies. All-item parcels were used for each and the residual variance of the indicator variables was freed. We used a situation-awareness global assessment test [18] during game 2 to assess situation awareness. The situation awareness questionnaire contained 5 questions related to the current game state. The situation awareness question items each contained a slider (min: 0%, max:100%) and asked participants to estimate: their current cooperation rate (1) and the hotdog (2,3) and lobster (4,5) reciprocity rates for each co-diner. The game interface was not available at this time. The situation awareness score was calculated by first summing up the errors from each of the 5 estimation questions and then inverting the scores based on the participant with the highest error, such that higher situation awareness scores are better. Finally, true expertise was measured just after the training with the training materials removed. This test was a set of eleven questions that relates to knowledge of the game and whether the participant possessed the ability to map from the current game state to the optimal decision:

(1) How much does a hotdog cost? (slider response)

[2]http://lesswrong.com/lw/7f2/prisoners_dilemma_tournament_results/

(2) *How much does a lobster cost? (slider response)*
(3) *What is the quality of a hotdog? (slider response)*
(4) *What is the quality of a lobster? (slider response)*
(5) *In a one-round Diner's Dilemma game (only one restaurant visit), you get the least amount of dining points when... (four options)*
(6) *In a one-round Diner's Dilemma game (only one restaurant visit), you get the most amount of dining points when... (four options)*
(7) *Which situation gets you more points? (two options)*
(8) *Which situation gets you more points? (two options)*
(9) *Suppose you know for sure that your co-diners reciprocate your Hotdog order 100% of the time and reciprocate your Lobster order 100% of the time. Which should you order for the rest of the game? (H/L)*
(10) *Suppose you know for sure that your co-diners reciprocate your Hotdog order 0% of the time and reciprocate your Lobster order 100% of the time. Which should you order for the rest of the game? (H/L)*
(11) *Suppose you know for sure that your co-diners reciprocate your Hotdog order 50% of the time and reciprocate your Lobster order 50% of the time. Which should you order for the rest of the game? (H/L)*

Hypotheses

Our analysis first considered the validity of the task and study design. Most importantly, the Dining Guru without errors should perform much better than the participants.

- H_1: The Dining Guru in the no error condition outperforms the participants on average.

Next, measured expertise and reported familiarity should have consequences for performance. To validate that self-reported familiarity quantifies the illusory superiority effect, reported familiarity should also negatively correlate with measured expertise, which in turn should positively correlate with performance. Thus:

- H_2: Self-reported familiarity predicts decreased performance.
- H_3: Measured expertise predicts increased performance.
- H_4: Self-reported familiarity predicts decreased expertise.

To answer the research questions, we used a combination of ANOVA and structural equation modeling (SEM) [55] . To address the first research question, we test the relationships among measured expertise, reported familiarity, adherence, and trust via SEM:

- H_5: Reported familiarity predicts decreased adherence and trust.
- H_6: Measured expertise predicts increased adherence and trust.

To address the second research question, we test for a general effect of explanation on adherence. To generate a more fine-grained picture of this effect, we break self-reported familiarity into quartiles and test where differences exist.

- H_7: Explanations increase adherence regardless of self-reported familiarity.

Third, we study the relationships between system variables and its effect on automation bias. We hypothesized that occurrence of automation bias may depend on the particular system design, for instance, making the system manual or presenting explanations may reduce automation bias. For this, automation, explanation, and error (as well as their interaction effects) are regressed onto situation awareness, adherence, and trust in the SEM for a total of 18 hypotheses.

To further investigate automation bias, we test the relationships between reported familiarity and measured expertise on situation awareness. Theories of illusory superiority and automation bias might lead us to hypothesize that over-confident users (who may be less likely to use the adviser) may have better situation awareness. Additionally, we hypothesized that explanations might have different effects on awareness at different levels of measured expertise and self-reported familiarity. To test this, we build an indicator product of the interactions effects between reported familiarity/explanation (which was 0 or 1) and measured expertise/explanation [43]. This produced the variables: $Familiarity \cdot Explanation$ and $Expertise \cdot Explanation$. Then, we test the following hypotheses:

- H_8: Self-reported familiarity predicts increased situation awareness.
- H_9: Situation Awareness predicts decreased adherence.
- H_{10}: $Expertise \cdot Explanation$ predicts increased situation awareness.
- H_{11}: $Familiarity \cdot Explanation$ predicts increased situation awareness.

For these final hypotheses, we also break expertise and familiarity up into quartiles and run an ANOVA to find the particular condition where situation awareness is being impacted (see Figure 5).

6 RESULTS

This section reports the results from the online experiment. Two SEMs are used to answer the research questions, the first to validate the task (Figure 3) and the second to build a model of participant data that can provide evidence for or against each individual hypothesis (Figure 4). For the uninitiated, an SEM is essentially a model of simultaneous multiple regressions that also allows for mediation analysis and latent factor modeling (these latter two make it the best choice for testing our hypotheses). Each circle in the reported SEMs indicate a latent factor (a factor built on multiple question items to better account for variance), each square indicates an observed variable, each arrow indicates a regression that

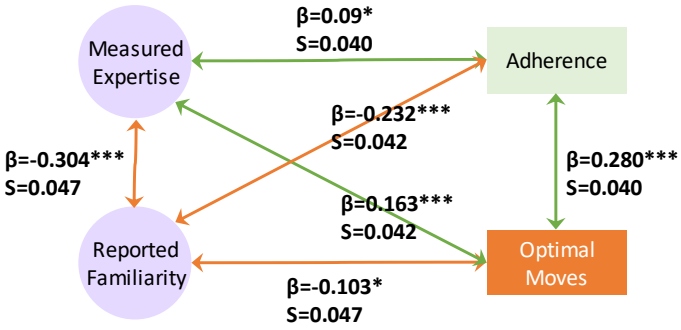

Figure 3: Correlations tested for validation of the task.

was fit with either β (normalized effect size in standard deviations) or B value for treatment variables (effect size observed when treatment is switched on). Co-variances are also reported (shown as bi-directional arrows) and the choice of regression or co-variance modeling was based on the relationship between variables (effect sizes significance values for these two choices are typically similar and the modeling choice between the two is a question of interpretation). Our SEM models were fit using lavaan[3] in R. A complete overview of SEM is given in [55]. Multiplicity control on our 29 hypotheses (11 explicit, 18 exploratory) was enforced using the Benjamini-Hochberg procedure with $Q = 0.05$ [3], which is recommended for SEM analysis [12]. Significance levels in this section are reported as follows: *** = $p < .001$, ** = $p < .01$, and * = $p < .05$.

867 participants attempted the study on Amazon Mechanical Turk (AMT) and eventually 551 completions were obtained (64% completion). Upon inspection of the pre-study data, we suspect that many of the dropouts were likely bots attempting to game our Qualtrics system, but failed when reaching the more complex game portion of the task. Participants that completed the study were paid $3.00 and spent an average of 11 minutes 24 seconds to complete the Diner's Dilemma training, and 15 minutes 30 seconds to complete all three games. Participants were between 18 and 70 years of age and 54% were male. Recruitment was opened up internationally to improve generalizability of results, but we required that workers to have 50 prior completed AMT tasks and be fluent in English. The data was checked for satisficing by examining input patterns (e.g., marking the same value for every question item) and those users were removed, resulting in a cleaned dataset of $N = 529$ participants. Attention was also controlled for by our CRT metric (it is not unreasonable to assume many users of such systems will be inattentive "in the wild," thus we did not remove these users).

[3]http://lavaan.ugent.be/

Task and Metric Validation

Simulations for the Dining Guru were run offline to establish its performance, which was compared with the participant data. Table 4 shows that, on average, participants performed slightly worse than the most error-prone version of the Dining Guru. This means that participants were on average better off when adhering to the Dining Guru's advice, regardless of error. It is also evidenced by the significance of the correlation between adherence and optimal moves shown in Figure 3. Thus we accept H_1.

Figure 3 shows correlations in participant data that explain relationships between measured expertise, reported familiarity, and the number of optimal moves. A significant negative correlation was found between reported familiarity and optimal moves ($\beta = -0.103*$) and a significant positive correlation was found for measured expertise ($\beta = 0.163***$). A large negative correlation was found between measured expertise and self-reported familiarity ($\beta = -0.304 * **$). Thus we accept H_2, H_3, and H_4.

Illusory Superiority

Next we test hypotheses related to illusory superiority. An SEM was built using the lavaan statistical package in R (Figure 4). Reported familiarity and measured expertise were used as regressands to explain adherence and trust. Significant effects were found for reported familiarity on trust ($\beta = 0.241 * **$) and adherence ($\beta = -0.289$). No direct effects were found for measured expertise, however, the model predicts that measured expertise does slightly increase adherence through a full mediation effect by situation awareness. This effect is also demonstrated by the weak correlation shown in Figure 3. Thus, we reject H_5, H_6.

No general effect of explanation was found on adherence. However, Figure 5 demonstrates that explanations did cause a significant increase in adherence for the participants who reported being very unfamiliar with the task (from about 30% adherence to about 40%). No effect is found on the upper quartiles, but a decreasing trend in adherence can be seen as self-reported familiarity increases. Thus we reject H_7.

Automation Bias

Next, we tested the effects and interactions of automation, error, and explanation to predict trust, adherence, and situation awareness. An effect of *Automation* · *Explanation* was found at the $p = 0.029$ level, but did not pass our multiplicity control. Error was found to cause reduced adherence ($B = -0.151 * *$), but this effect does not hold for the automated system ($B = 0.173 * **$). Automation was also found to cause decreased situation awareness ($B = -0.100*$). This exploratory analysis confirmed automation bias, but also indicates that explanations may not be an effective remedy.

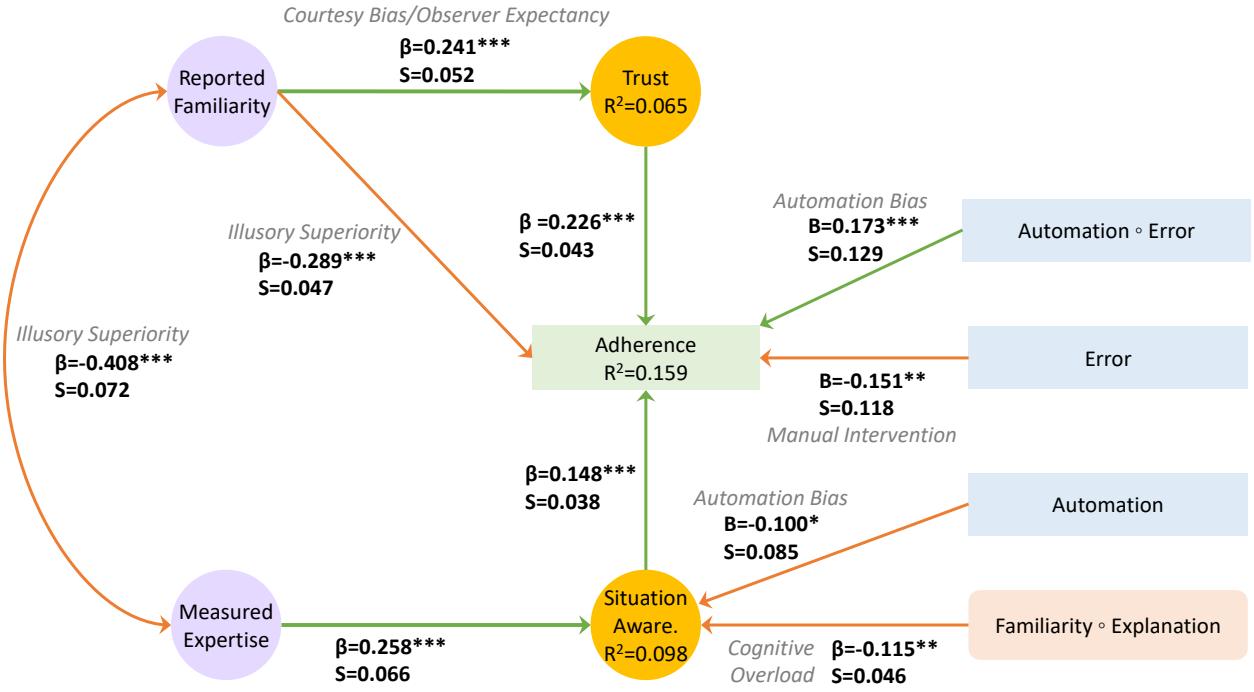

Figure 4: An SEM built to test hypotheses: unidirectional arrows indicate regression, bidirectional arrows indicate covariance; red arrows indicate a negative effect, green arrows indicate a positive effect; latent factors were scaled so β values indicate effect sizes in units of standard deviations; system characteristics (which were 0 or 1) were not scaled, so B values are reported. Standard error (S) is given. Model fit: $N = 529$ with 35 free parameters = 15.1 participants per free parameter, $RMSEA = 0.028$ $(CI : [0.017, 0.038])$, $TLI = 0.979, CFI > 0.983$ over null baseline model, $\chi^2(112) = 158.072$. Text in grey indicates rationale.

Game #	Optimal Choice	# Rounds	DG/No Error	DG/Weak Error	DG/Error	Participants
1	LOBSTER	55	0.92	0.75	0.63	0.70
2	HOTDOG	60	0.65	0.62	0.56	0.46
3	LOBSTER	58	0.79	0.69	0.60	0.62
All		17	0.78	0.68	**0.60**	**0.59**

Table 4: Performance of the Dining Guru (DG) across all games compared to participants. The ratio of optimal moves made by the error-free Dining Guru (DG/No Error), weak-error Dining Guru (DG/Weak Error), full-error Dining Guru (DG/Error), and participants are given. DG/Error performed as well as the participants on average.

Finally, we test hypotheses about the relationship between illusory superiority and automation bias. It was found that measured expertise was a significant predictor of situation awareness ($\beta = 0.258 * **$), but no effect was found between reported familiarity and situation awareness. However, measured expertise has a controlling effect on reported familiarity – if measured expertise is removed from the model, then reported familiarity becomes a significant predictor of situation awareness (this is also another good indicator that our factor measurements are valid). The indicator products of $Expertise \cdot Explanation$ and $Familiarity \cdot Explanation$ were also regressed onto situation awareness. A significant effect was found for $Familiarity \cdot Explanation$ ($\beta = -0.115 * *$). This can also be seen in Figure 5, where it is evident that explanations caused decreased awareness in the fourth quartile of self-reported familiarity.

7 DISCUSSION

We derived three key takeaways from the analysis, which are discussed in detail in this section:

(1) Reported trust and behavioral adherence have a weak relationship, therefore we encourage more open-ended methodologies in intelligent systems research.
(2) Rational explanations are only effective on users that report being very unfamiliar with a task – regardless of their actual competency level.

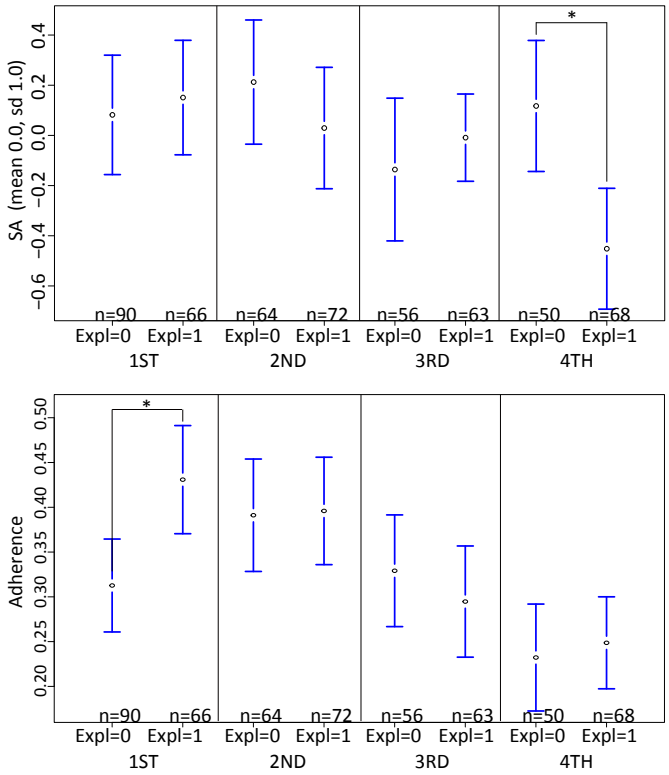

Figure 5: Changes in situation awareness (SA) and adherence caused by explanation (Expl.), broken up by quartile (1st, 2nd, 3rd, 4th) of self-reported familiarity. Error bars are one standard deviation. Explanations cause a significant decrease in SA in the 4th quartile of familiarity and are only effective on users who report very low familiarity with the task.

(3) There is a danger in showing explanations to self-confident users in that situation awareness might be negatively impacted – this can be mitigated by requiring interaction with an agent.
(4) Requiring interaction (more "manual" systems) is not a cure for poor adherence due to overconfidence – this suggests that non-rational methods, such as appealing to emotion, may be the only avenue to accommodate the overconfident.

Initially, we had hypothesized that over-confident users would not use or trust the Dining Guru. While we were partially correct, it was surprising that these participants used the Dining Guru less while at the same time reporting to trust it more. There are a few possible explanations for this, the simplest being that users simply lied about trusting the Dining Guru because they knew they were being evaluated (a type of courtesy bias[4] or observer-expectancy effect [26]). A more complicated theory of trust in our data is that many participants believed the Dining Guru gave good advice, but simply believed they could do better, due to illusory superiority. This is because people are very likely to rate themselves in the top 15% of performers and thus expected to be able to do better than the Dining Guru. Dunning-Kruger's "double burden" effect also shows up in our participant data: while we had hypothesized that users that adhered less would show more situation awareness, the opposite showed itself to be true. This might be because users of higher cognitive ability consequently displayed higher expertise in the game, were better able to keep track of the game state, and thus were more likely to recognize that the Dining Guru was effective and agree with it. We also found that the relationship between trust and adherence was not one-to-one – this finding shows limitations in experiment methodologies that only account for self-reported metrics, which are very common in recommender systems research. This echoes the observation that quantification of the Dunning-Kruger effect on adherence was possible in this study because participants were given two alternative methods for solving the task – a lack of participant freedom in modern intelligent assistant studies may have contributed to this effect not being detected until now.

In many ways, explanations were not an effective method for encouraging use of the Dining Guru, improving trust, or preventing automation bias. Explanations were only effective on the participants that could admit they were not familiar with the task almost at all (Figure 5) and its effects quickly drop off as over-confidence increases. This builds on the work by Arnold et al. [1], who found that self-reported novices were more likely to adhere to advice. Moreover, explanations appear to exacerbate automation bias for the most over-confident users, as evidenced by Figure 5 (top, right) and $Familiarity \cdot Explanation$. The most likely explanation for this is information overload [21], also referred to as cognitive overload. Although it is not statistically reported here, a cognitive load metric was taken as a control during the study and analysis showed a negative correlation with situation awareness, supporting this conclusion. Moreover, our analysis has established reported task familiarity as a quantifier of illusory superiority, which means that it may indicate relatively lower cognitive ability. This could explain susceptibility to the cognitive overload state.

Effects related to automation bias [13] were reproduced here. Unsurprisingly, automation predicted slightly decreased situation awareness. The effect size here may not be as large

[4]https://www.theguardian.com/politics/2015/may/08/election-2015-how-shy-tories-confounded-polls-cameron-victory

as in other tasks (e.g., auto-piloting[5]), since the automated version of the Dining Guru did not "take control" of the decision making. When the manual version of the Dining Guru was presented, our participants demonstrated the ability to recognize the system was making mistakes and intervene. This was not the case when the system was automated (see Figure 4, right side - automation bias and manual intervention effects). To investigate this further, we examined the number of times the Dining Guru was solicited for advice based on automation and reported familiarity. As expected, users with higher reported task familiarity accessed the Dining Guru less, but automation caused an increase in solicitation, likely because it was easier to use. This increase in usability might explain the automation bias effect.

The results from this study pose a problem for the deployment of intelligent assistants and interfaces. As previously discussed, game theory indicates that the iterated prisoner's dilemma and its derivatives model many real-world decision-making situations and we have attempted to mirror that generality with our "free choice" task design. In the Diner's Dilemma as in many other situations, knowledge automation has an enormous potential to benefit those with lower cognitive ability, since it gives them access to cognitive resources and domain knowledge that they previously lacked or were hard-pressed to access. However, if many of these users are unable to identify the situations in which they need to release the reins (Kruger's "double burden"), intelligent assistants would be of little value. For example, consider a situation where a self-confident individual enters the gym for the first time with the intent of "figuring it out" rather than taking advice from a strength or weight-loss coach. This is very similar to our task setup in that multiple decision-making iterations are conducted and at each point the user has the option to consult with an intelligent assistant or coach – the only difference being that the Diner's Dilemma is a binary choice task while training options are multiple. In this case, the results from this study indicate that a fitness recommender system might be ill-advised to recommend exercises, programs, or even nutrition. Moreover, explanations that rationally explain the benefits of different nutrition or training programs might not do anything to sway this gym-goer (instead, the resulting cognitive load might even cause the system to be uninstalled for its "poor user interface"). The consequence is that our gym-goer might end up wasting a lot of time in the gym, with no obvious penalty from their perspective, because an implicit, rather than explicit, cost is being paid. This type of implicit cost situation extends to many other applications, including movie recommendation (where alternative options might never be viewed), driving, career choice, and military spending[6].

What can be done to sway over-confident users into taking system advice? At this time, we cannot definitively answer that question. A final exploratory analysis was performed to determine if the over-confident users were more likely to use the Dining Guru if the manual version was presented. No significant effect was found and, surprisingly, the trend had a negative direction. Dunning-Kruger's "double burden" may be an intractable problem for system design: for these users, an automated system might be perceived to be incompetent but a manual system might be too hard or bothersome to use. Can these users be accommodated some other way? Perhaps a way forward is by exploiting the user's psychology in the form of continuous status updates on the performance of his or her peers. For instance, our intelligent assistant could notify a gym-goer that his fitness level is below average for people at his experience level, although this may only work on people who are motivated to avoid negative consequences [20]. Another option is to appeal to emotions [22], rather than rationality [5, 44]. Finally, the risk of the domain may complicate how self-confident users behave. Here, the cost incurred to the participant based on their decisions was likely perceived to be low, but if the risk was high, behavior may vary. A follow-up study will experiment with the framing of the intelligent assistant in terms of personality, competency, and authority, as well as the social information it displays and the level of risk incurred by each choice through the use of monetary reward incentives.

8 CONCLUSION

In conclusion, we conducted a crowd-sourced user study (N=529) on participants playing the Diner's Dilemma game with help from a virtual agent – the Dining Guru. An analysis was conducted to understand the implications of illusory superiority for use of intelligent assistants. Participants that considered themselves very familiar with the task domain reported higher than average trust in the Dining Guru but took its advice less often. Presenting explanations was ineffective and in some cases led these users to automation bias. The results from this study suggest that rational explanations are only effective at lower levels of self-assessed knowledge. A new strategy for improving trust and adherence may need to be designed for over-confident users, such as appealing to emotion or adapting systems based on personality.

[5]http://www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash

[6]https://www.sipri.org/commentary/blog/2016/opportunity-cost-world-military-spending

REFERENCES

[1] Vicky Arnold, Nicole Clark, Philip A Collier, Stewart A Leech, and Steve G Sutton. 2006. The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *Mis Quarterly* (2006), 79–97.

[2] Robert Axelrod and Robert M Axelrod. 1984. *The evolution of cooperation*. Vol. 5145. Basic Books (AZ).

[3] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* (1995), 289–300.

[4] Shlomo Berkovsky, Ronnie Taib, and Dan Conway. 2017. How to Recommend?: User Trust Factors in Movie Recommender Systems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 287–300.

[5] John Boddy, Annabel Carver, and Kevin Rowley. 1986. Effects of positive and negative verbal reinforcement on performance as a function of extraversion-introversion: Some tests of Gray's theory. *Personality and Individual Differences* 7, 1 (1986), 81–88.

[6] Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin* 56, 2 (1959), 81.

[7] Jennifer A Chatman and Sigal G Barsade. 1995. Personality, organizational culture, and cooperation: Evidence from a business simulation. *Administrative Science Quarterly* (1995), 423–443.

[8] Jessie YC Chen and Michael J Barnes. 2014. Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems* 44, 1 (2014), 13–29.

[9] Jessie Y Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael Barnes. 2014. *Situation Awareness-Based Agent Transparency*. Technical Report. DTIC Document.

[10] Yu-Hui Chen and Stuart Barnes. 2007. Initial trust and online buyer behaviour. *Industrial management & data systems* 107, 1 (2007), 21–36.

[11] Ahyoung Choi, Celso M de Melo, Peter Khooshabeh, Woontack Woo, and Jonathan Gratch. 2015. Physiological evidence for a dual process model of the social effects of emotion in computers. *International Journal of Human-Computer Studies* 74 (2015), 41–53.

[12] Robert A Cribbie. 2007. Multiplicity control in structural equation modeling. *Structural Equation Modeling* 14, 1 (2007), 98–112.

[13] Mary Cummings. 2004. Automation bias in intelligent time critical decision support systems. In *AIAA 1st Intelligent Systems Technical Conference*. 6313.

[14] Janet E Davidson and CL Downing. 2000. Contemporary models of intelligence. *Handbook of intelligence* (2000), 34–49.

[15] Robyn M Dawes and Richard H Thaler. 1988. Anomalies: cooperation. *The Journal of Economic Perspectives* (1988), 187–197.

[16] David Dunning. 2011. 5 The Dunning-Kruger Effect: On Being Ignorant of One's Own Ignorance. *Advances in experimental social psychology* 44 (2011), 247.

[17] Mica R Endsley. 1988. Situation awareness global assessment technique (SAGAT). In *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National*. IEEE, 789–795.

[18] Mica R Endsley. 2000. Direct measurement of situation awareness: Validity and use of SAGAT. *Situation awareness analysis and measurement* 10 (2000).

[19] Mica R Endsley and Daniel J Garland. 2000. *Situation awareness analysis and measurement*. CRC Press.

[20] Renee Engeln-Maddox. 2005. Cognitive responses to idealized media images of women: The relationship of social comparison and critical processing to body image disturbance in college women. *Journal of Social and Clinical Psychology* 24, 8 (2005), 1114–1138.

[21] Martin J Eppler and Jeanne Mengis. 2004. The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The information society* 20, 5 (2004), 325–344.

[22] Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity* 2002, December (2002), 5.

[23] William H Gladstones, Michael A Regan, and Robert B Lee. 1989. Division of attention: The single-channel hypothesis revisited. *The Quarterly Journal of Experimental Psychology* 41, 1 (1989), 1–17.

[24] Natalie S Glance and Bernardo A Huberman. 1994. The dynamics of social dilemmas. *Scientific American* 270, 3 (1994), 76–81.

[25] Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist* 48, 1 (1993), 26.

[26] E Bruce Goldstein. 2014. *Cognitive psychology: Connecting mind, research and everyday experience*. Nelson Education.

[27] C Gonzalez, N Ben-Asher, JM Martin, and V Dutt. 2013. Emergence of cooperation with increased information: Explaining the process with instance-based learning models. *Unpublished manuscript under review* (2013).

[28] Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.

[29] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).

[30] Jason L Harman, John O'Donovan, Tarek Abdelzaher, and Cleotilde Gonzalez. 2014. Dynamics of human trust in recommender systems. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 305–308.

[31] Morten Hertzum, Hans HK Andersen, Verner Andersen, and Camilla B Hansen. 2002. Trust in information sources: seeking information from people, documents, and virtual agents. *Interacting with computers* 14, 5 (2002), 575–599.

[32] David Hitchcock and Bart Verheij. 2006. *Arguing on the Toulmin model*. Springer.

[33] Vera Hoorens. 1993. Self-enhancement and superiority biases in social comparison. *European review of social psychology* 4, 1 (1993), 113–139.

[34] Jason D Johnson, Julian Sanchez, Arthur D Fisk, and Wendy A Rogers. 2004. Type of automation failure: The effects on trust and reliance in automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48. SAGE Publications Sage CA: Los Angeles, CA, 2163–2167.

[35] John Kagel and Peter McGee. 2014. Personality and cooperation in finitely repeated prisonerâĂŹs dilemma games. *Economics Letters* 124, 2 (2014), 274–277.

[36] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 43–50.

[37] Bart P Knijnenburg and Alfred Kobsa. 2013. Making decisions about privacy: information disclosure in context-aware recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 3 (2013), 20.

[38] Sherrie YX Komiak and Izak Benbasat. 2006. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS quarterly* (2006), 941–960.

[39] Yehuda Koren and Robert Bell. 2015. Advances in collaborative filtering. In *Recommender systems handbook*. Springer, 77–118.

[40] Marios Koufaris and William Hampton-Sosa. 2002. Customer trust online: examining the role of the experience with the Web-site. *Department of Statistics and Computer Information Systems Working Paper Series, Zicklin School of Business, Baruch College, New York* (2002).

[41] Justin Kruger. 1999. Lake Wobegon be gone! The" below-average effect" and the egocentric nature of comparative ability judgments. *Journal of personality and social psychology* 77, 2 (1999), 221.

[42] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[43] Guan-Chyun Lin, Zhonglin Wen, Herbert W Marsh, and Huey-Shyan Lin. 2010. Structural equation models of latent interactions: Clarification of orthogonalizing and double-mean-centering strategies. *Structural Equation Modeling* 17, 3 (2010), 374–391.

[44] Charles G Lord, Lee Ross, and Mark R Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology* 37, 11 (1979), 2098.

[45] Jolie M Martin, Ion Juvina, Christian Lebiere, and Cleotilde Gonzalez. 2011. The effects of individual and context on aggression in repeated social interaction. In *Engineering Psychology and Cognitive Ergonomics*. Springer, 442–451.

[46] D Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)* 2, 2 (2011), 12.

[47] Harrison McKnight, Michelle Carter, and Paul Clay. 2009. Trust in technology: development of a set of constructs and measures. *Digit 2009 Proceedings* (2009), 10.

[48] Kimberly Merritt, D Smith, and JCD Renzo. 2005. An investigation of self-reported computer literacy: Is it reliable. *Issues in Information Systems* 6, 1 (2005), 289–295.

[49] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (2017), 393–444.

[50] John O'Donovan and Barry Smyth. 2005. Trust in recommender systems. In *Proceedings of the 10th international conference on Intelligent user interfaces*. ACM, 167–174.

[51] Anatol Rapoport and Albert M Chammah. 1965. *Prisoner's dilemma: A study in conflict and cooperation*. Vol. 165. University of Michigan press.

[52] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 141–148.

[53] James Schaffer, John OâĂŹDonovan, Laura Marusich, Michael Yu, Cleotilde Gonzalez, and Tobias Höllerer. 2018. A study of dynamic information display and decision-making in abstract trust games. *International Journal of Human-Computer Studies* 113 (2018), 1–14.

[54] Mohsen Tavakol and Reg Dennick. 2011. Making sense of Cronbach's alpha. *International journal of medical education* 2 (2011), 53.

[55] Jodie B Ullman and Peter M Bentler. 2003. *Structural equation modeling*. Wiley Online Library.

[56] Weiquan Wang and Izak Benbasat. 2007. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23, 4 (2007), 217–246.

[57] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, 307–317.