Knowledge Complacency and Decision Support Systems

Sebastian S. Rodriguez
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA
srodri44@illinois.edu

James Austin Schaffer
Battlefield Information Processing Branch
US Army Research Laboratory West
Playa Vista, CA, USA
james.a.schaffer20.civ@mail.mil

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

Tobias Höllerer
Department of Computer Science
University of California, Santa Barbara
Santa Barbara, CA, USA
holl@cs.ucsb.edu

***Abstract*—Decision support systems (DSS), which are often based on complex statistical, machine learning, and AI models, have increasingly become a core part of data analytics and sensemaking processes. Automation complacency – a state characterized by over-trust in intelligent systems – has the potential to result in catastrophic performance failure. An under-investigated factor in automation complacency research is the effect that DSS might have on human learning of domain concepts. In this paper, we perform a comparative analysis of two studies of users interacting with decision aids to understand how knowledge retention is affected by the competence and presentation of a DSS. Our results indicate that while humans have the opportunity to learn and internalize domain concepts while being supported by a DSS, features that make the DSS appear more competent, persuasive, or customizable may lead a user to form incorrect beliefs about a domain.**

Keywords*—decision support systems, human-machine cognition, automation complacency, recommender systems, interfaces, user studies, artificial intelligence, virtual agents

I. Introduction

When performing information search, requirements can only be met if analysts are able to internalize the results as knowledge. Despite this, complex information systems, such as decision support systems (DSS), have become increasingly sophisticated, which consequently makes them difficult to comprehend and predict. Designers of DSS have responded by increasing transparency (via explanations) [1], [2] and customizability/feedback (via control parameters) [3]. Although these system choices have been shown to increase trustworthiness and user adoption [4], it may also cause states of over-trusting, which may lead to automation complacency [5]. Additionally, as cognitive tasks are increasingly automated by DSS, users may lose the opportunity to exercise tacit and procedural knowledge. Despite this, computer systems are not yet flexible, accurate, or intelligent enough to robustly handle all unforeseen situations. Therefore, humans must continue to remain "in the loop" [6], taking the role of automation supervisor and decision maker [7].

DSS have evolved to decrease human mental effort and improve the amount of data that can be incorporated into the decision making process. Complex information tools now automatically summarize data and provide recommendations for decisions. These systems thus provide easy access to stored procedural knowledge and benefit from expertise possibly not known at the time of use. Examples of these systems include path-finding algorithms for automobile navigation [8] and collaborative filtering for product recommendations [9]. The impact of the use of DSS on bias and complacency [10], performance degradation [5], and situation awareness [11] have been well documented. However, it is still not clear how these systems impact short-term knowledge retention of their users and longitudinal learning. For instance, drivers that navigate by memorizing a route ahead of time (and repeating this action) may retain knowledge of the route for much longer than a GPS-assisted driver. Users that retain procedural knowledge of their actions would be more likely to perform better in situations when the automated aid is unavailable.

In this paper, we present an analysis of two experiments that demonstrate how DSS can affect a user's domain knowledge retention under certain levels of explanation, control, and error from a DSS. Explanation in decision aids is the visual indication of reasoning logic from the decision aid. Control is the amount of input that the systems allows or needs from the user. Error is the amount of noise and inaccuracies that originates from the data, as to simulate real-world use. The first study is a typical study of item recommendation (initially reported by Schaffer et al. [12]) wherein users performed an open ended choice task with support from a collaborative filtering recommender system. The second study is of the Diner's Dilemma, similar to Onal et al. [13] and Schaffer et al. [14], except in the present study an automated agent provides recommendations to the player each round. We then establish the following 4 hypotheses:

978-1-5386-9599-9/19/$31.00 ©2019 IEEE

Fig. 1. Interface for the Movie study, showing the browser tool (left, blue) and the recommender system (right, red). Participants found five to seven movies they were interested in and added them to their watchlist (middle, green). The recommender functioned off of the profile data (middle, yellow), which was specified beforehand.

- H_1: Inaccurate DSS (i.e., error-prone) lead to increased knowledge over time.
- H_2: Explanations from DSS prevents knowledge from decreasing over time.
- H_3: Control over DSS leads to increased knowledge over time.
- H_4: Control over DSS along with explanations leads to increased knowledge over time.

Our comparative analysis of both studies suggests that knowledge retention is not only affected by the competence of the information system but also whether it provides explanation or allows user interaction.

II. Related Work

Here we review relevant work in automation and knowledge.

A. Issues in Automation

Users of computer systems often behave like supervisors of complex automated algorithms, only intervening when decisions need to be made or when automation fails. This human-machine setup has many demonstrated benefits, for instance, by the overall reduction in airline crashes due to reliable autopilot systems. However, these systems can allow for catastrophic failures, as evidenced by the crash of Northwest Airlines flight MD-80, the 1989 US Air B-737 crash, and the 1983 Korean Airlines navigational failure [11]. Different types of issues present in the pilot-autopilot interface contributed to the cause of each one of these crashes. Research has investigated the issues inherent in this human-machine setup, including automation complacency, bias, and performance degradation.

Automation complacency refers to a condition where human operators of an automated system "over-trust" the machine performance, which results in a reduced frequency of checks to verify the machine is functioning properly [5], and cannot be prevented by simple training or instructions [15]. Recent research suggests that training protocols can reduce task performance caused by automation complacency [10]. Additionally, operators that are at-risk for automation complacent states can be detected using simple questionnaires [16].

Automation bias occurs in the use of decision support systems when users take recommendations without seeking out additional sources to verify the effectiveness of the recommendation [17]. This can easily lead to undesirable situations when systems make catastrophic errors. Additionally, interacting with complex algorithms can significantly alter user beliefs about data spaces [18].

Embodied automated systems (typically referred to as vir-

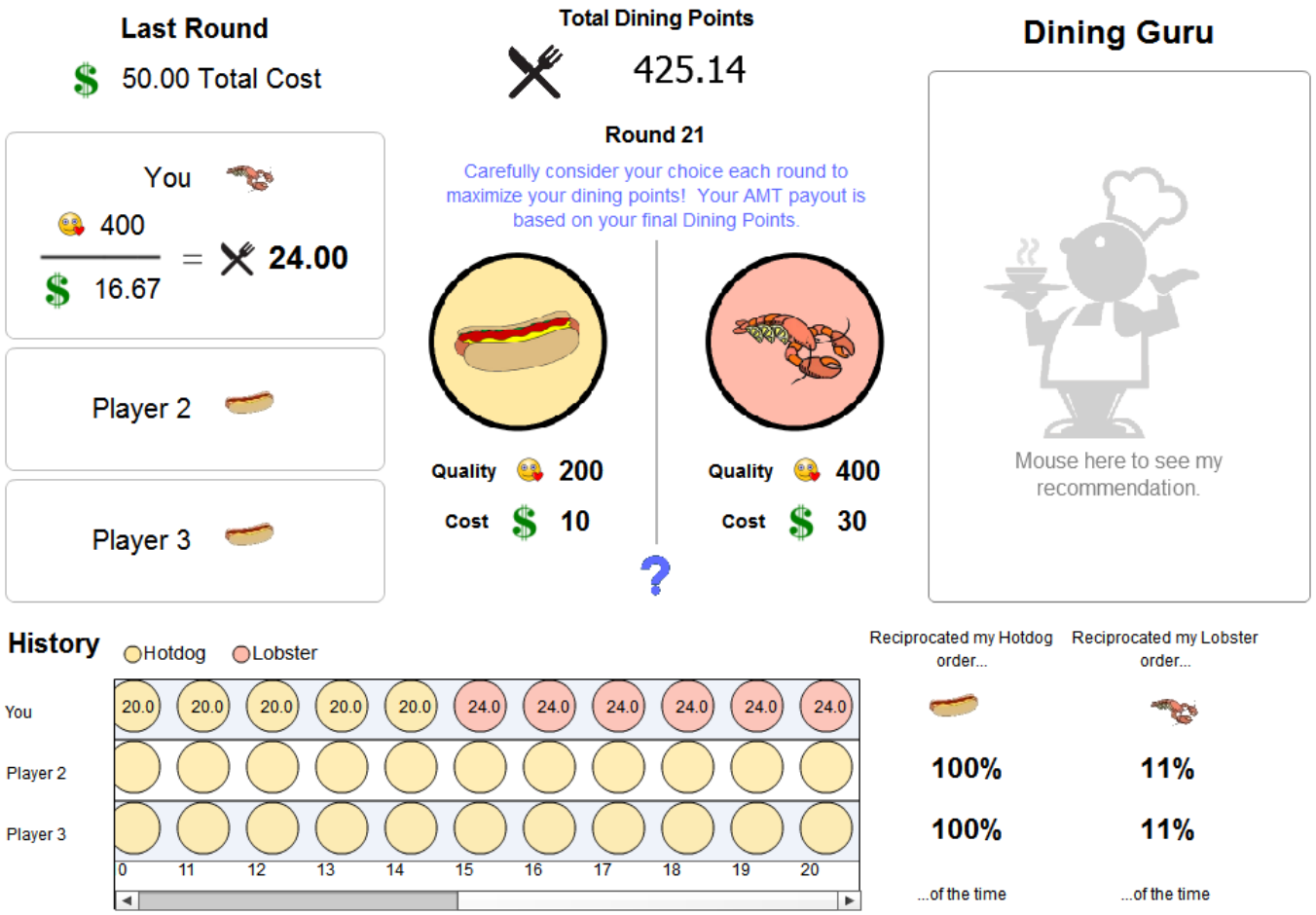

Fig. 2. Interface for the Diner study, showing the core game information (center, left), the history tool (bottom), and the Dining Guru (right side). Participants could seek recommendations by mousing over the Dining Guru or alternatively use the history panel to devise strategies.

tual agents) are known to have stronger personified features [19] than their non-embodied "artifact-like" counterparts [20]. It has been demonstrated that people can form an opinion of a virtual agent within the first few seconds of interaction and become more conscientious about their behaviors [21], reacting as if the agents were real people. Despite this, trust relationships with non-embodied agents – especially recommender systems – continue to be studied [3], perhaps because the alternative remains expensive and their performance is considered an open question [22], [23]. Embodied agents may therefore pose a larger problem for complacency than their counterparts.

B. Knowledge

When analyzing new information, humans accumulate insights which build up into broader knowledge of the domain [24]. While the research community do not agree on an exact definition, Saraiya et al. define insight as "an individual observation about the data by the participant, or a 'unit of discovery"' [25]. North et al. offers a compelling characterization of insight [26]: insight is complex, deep, qualitative, unexpected, and relevant. These characteristics can serve as guidelines with which to design benchmark questionnaires to assess insight in specific domains. We take a similar approach by using testing methodologies to assess knowledge, wherein each question on the test amounts to a single insight.

The prior knowledge of users also impact whether recommendations and their explanations should be present. To illustrate, Arnold et al. showed that the direction of explanations depends on the expertise of the user, where novices prefer feedforward rather than feedback [2]. Additionally, DSS have been shown to both reduce or aggravate biases in users [27].

Knowledge complacency triggered by DSS has the capacity to interfere with longitudinal insight retention – spaced repetition [28] and memory retrieval [29] are important for long term knowledge retention. By providing a "decision aid," users might become deprived of opportunities to recall information that may be critical for decision making. Over repeated sessions, users that rely on DSS may gradually see a decline in the number of domain-related insights that can be recalled. For instance, repeatedly using a GPS system to navigate to a frequented destination with a complex set of turns may deny the operator the ability to internalize the route.

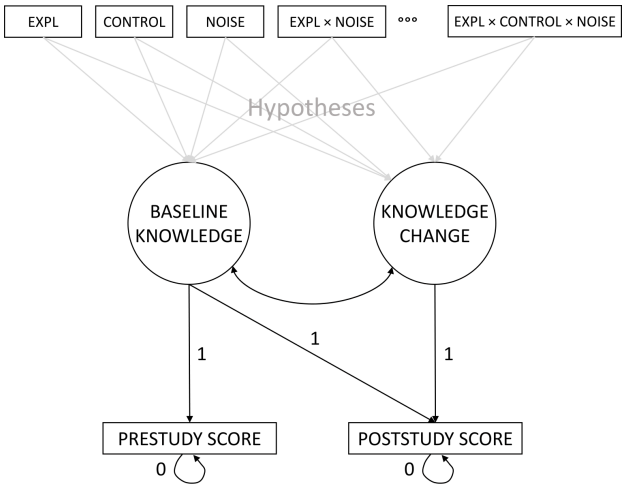

Fig. 3. The Raykov change score model that was specified to test hypotheses.

To our best knowledge, there is little work that examines how the use of complex automated algorithms affect user accumulation of domain knowledge and understanding of data spaces (with Schaffer et al. being an exception [18]). Difficulties in defining adequate benchmarks that successfully measure domain knowledge in the laboratory remains a significant research challenge.

III. Methodology

In this section, we describe the methodology of each study in more detail. In both studies, we manipulated the level of explanation, user control, and system error from a DSS designed for the task. Methodology for each study was kept as similar as possible to enhance generalization. A key design point of each study was to give the participant the ability to complete the task while ignoring recommendations from the agent – this resulted in the design of the browser tool for the Movie study (Figure 1, left, blue) and the history tool for the Diner study (Figure 2, bottom, panel with yellow and red circles). Variables were manipulated in a between-subjects design, affecting the recommendation interface the participants used for each study.

In both studies, a knowledge test was given at two points: once at the outset of the task and once at the conclusion of the task. These knowledge tests consisted of 8 (Movie) or 11 (Diner) questions that queried domain knowledge. In the case of the Movie recommendation study, the questions were chosen such that someone who was experienced at searching movie databases (e.g., IMDb [30]) would be able to score higher than the average person. In the Diner study, the questions were chosen such that a game expert would score highly, and the questions were for the most part replicated from [13], wherein it was shown that performance on this test correlated with high performance in the game.

To analyze the predictors of knowledge change, we chose to use a Raykov change model (Figure 3) [31]. This test was used due to the assumption of non-normality in the test data and ease of execution using lavaan [32] in R [33]. A Raykov model is a particular pathway model that can handle mixed study designs. In the model, a special baseline and change factor are specified based on the two-wave test. Then, each predictor of change is regressed onto both the baseline and the change factor. This results in estimates for the intercept and slope of the predictor, respectively.

A. Movie Recommendation Study

This experiment was designed to simulate users selecting a movie to watch using available information tools online. Subjects were tasked with finding movies for their "watchlist", with a time limit of 12 minutes to select 7 movies. Subjects were provided with a simple data browsing tool (similar to what is commonly available, such as Rotten Tomatoes, IMDb, etc.) and a complex recommendation algorithm: collaborative filtering with Herlocker damping. Both the movie browser and recommender used identical interfaces (Figure 1). Subjects could freely interact with either tool.

The knowledge test was designed to gauge the user's knowledge of movie metadata. High scores would be indicative of a good internalization of the movie metadata domain. IMDb records were used as ground truth for the questions. 8 questions were administered, as follows:

1) Online, which genre has the highest current average audience rating?
2) Online, which of these genres tends to be the most common among the movies with the highest average audience rating?
3) Online, which of these genres has the highest current popularity?
4) Generally, which of these genres has the most titles released, for all time periods?
5) Online, which of these decades has the highest current average audience rating?
6) How many movies have an average audience rating great than 9/10?
7) Popular movies tend to have an average rating that is lower|average|higher?
8) Movies with an average rating of 9/10 or higher tend to have fewer|average|more votes?

Explanations were text-based and tied together the subject's movie profile with recommendations. Control features allowed subjects to customize what the recommender showed by filtering movie metadata. Error was added per condition ("no error"/"error"), where random noise was added to the collaborative filtering algorithm, which slightly modified the predicted ratings for subjects and shuffled the list of recommendations.

B. Diner's Dilemma Study

This experiment examined the iterated "prisoner's dilemma", which has been frequently used to model real-world scenarios. We created a web game called "Diner's Dilemma", a theme variation on the prisoner's dilemma. Subjects visit a restaurant with two of their friends, and

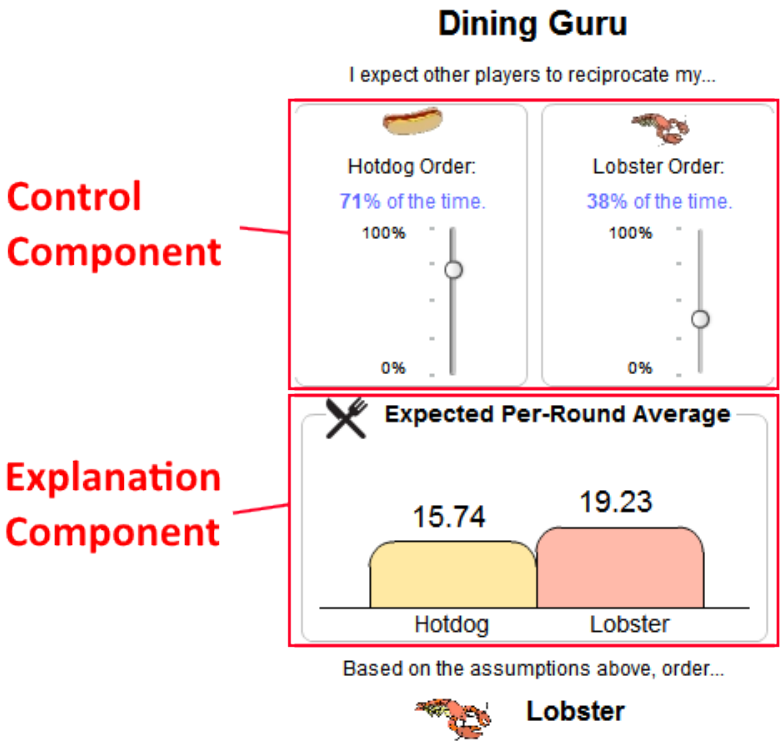

Fig. 4. Dining Guru components varied based on treatment. The control component and explanation component are shown in red. In the baseline treatment, the Dining Guru would only recommend the item (bottom).

must choose between the inexpensive dish (Hotdog), which has low nutritional value, or the expensive dish (Lobster), with high nutritional value. The diners agree to split the bill equally no matter what is ordered. As the game progresses, subjects must form strategies of cooperation or defection to maximize the points they score in each round, which is the nutritional value divided by the price they paid. The nuances of the iterated prisoner's dilemma have been well studied [34]. Moreover, a detailed description of the properties of the Diner's Dilemma game can be found in [14].

While playing, subjects were provided with a simple visualization of choices made and points earned by the group in previous rounds. Subjects also had the choice of taking recommendations from the "Dining Guru" (Fig. 2), which analyzed the group's behaviour and suggested a selection based on which item was expected to result in the most number of points. Unlike the movie recommendation study, users accessed recommendations on demand by mousing over the Dining Guru panel (see Figure 2). Users were trained in the interface panel and the Dining Guru's purpose before the game started.

Three games of Diner's Dilemma were played, consisting of approximately 60 rounds each. Simulated co-diner strategies (variations on tit-for-tat) changed between each game. The optimal strategy for each game was either to always cooperate or always defect, based on the current co-diner strategy.

Explanations were both visual and text-based, and connected the current recommendation to past co-diner behaviour. Control features allowed subjects to provide input into the Dining Guru which allowed them to freely explore the different cooperation/defection strategies encountered during the game (as seen in Figure 4). Due to the nature of the game, we added error at a higher level of granularity ("no error"/"low error"/"high error"). The error came in the form of random noise, which was added to recommendations between rounds to prevent the user from easily detecting them.

The knowledge test was designed to gauge the user's knowledge of the game. High scores would be indicative of a good internalization of the game's procedure. 11 questions were administered, as follows:

1) How much does a Hotdog cost? (slider response)
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? (Hotdog/Lobster)

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? (Hotdog/Lobster)
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? (Hotdog/Lobster)

IV. Results

In total, 1,055 participant records were used in the analysis. For the Movie study, we recruited more than 526 participants between 18 and 71 years of age (M = 35 years, SD = 11 years, 45% male) via Amazon Mechanical Turk. Participants spent between 25 and 60 minutes performing the task, and were compensated $1.50. Participant data was carefully curated for satisficing, resulting in 526 complete records. For the Diner study, we recruited more than 529 participants between 18 and 70 years of age (M = 34 years, SD = 10 years, 54% male) via Amazon Mechanical Turk. Participants spent between 30 and 50 minutes playing the game, and were compensated $3.00. Participant data was carefully curated for satisficing, resulting in 529 complete records.

Coefficient estimates and significances of both Raykov change models can be found in Table I, and solutions to the model can be found in Table II (this is derived by plugging the values of explanation, control, and error into the model for each specific treatment, e.g. $explanation = 0, control = 1, error = 1$).

A. Movie Recommendation Study

A McNemar's Chi-squared test with continuity correction revealed that knowledge test performance significantly differed after treatment ($X^2(1, 4208) = 10.263$, p = 0.0013, ϕ = 0.05, odds ratio is 10.3).

The Raykov change model reveals that explanation, control, and error all caused incorrect beliefs to be formed. Users in the "control only" condition scored less than half a standard deviation lower in the final knowledge test (solution = -0.527). This was somewhat mitigated with the presence of explanations (solution = -0.276) and error (solution = -0.347). "Error only" led to incorrect beliefs (solution = -0.391). Explanations, control, and error have significant interactions which led to incorrect beliefs (solution = -0.309).

B. Diner's Dilemma Study

A McNemar's Chi-squared test with continuity correction revealed that knowledge test performance significantly differed after treatment ($X^2(1, 5819) = 37.081$, p <0.001, ϕ = 0.08, odds ratio is 11.3).

The Raykov change model reveals that explanation, control, and error all caused incorrect beliefs to be formed. Lowest scores in the knowledge test originate from the "control only" (solution = -0.563), "explanation, control, and error" (solution = -0.576), and "explanation and error" (solution = -0.614) conditions.

C. Hypothesis Testing

H_1: Inaccurate DSS (i.e., error-prone) lead to increased knowledge over time. In both studies, the Raykov change model predicts that error decreases the subject's score in the knowledge test after the task (Movie B = -0.391, Diner B = -0.477). The coefficient was found to be significant in both the Movie study (p = 0.022) and the Diner study (p = 0.0025), thus we reject H_1.

H_2: Explanations from DSS prevents knowledge from decreasing over time. In both studies, the Raykov change model predicts that the presence of explanations decreases the subject's score in the knowledge test after the task (Movie B = -0.298, Diner B = -0.464). The coefficient was found to be marginally significant in the Movie study (p = 0.088), and significant in the Diner study (p = 0.015), thus we reject H_2.

H_3: Control over DSS leads to increased knowledge over time. In both studies, the Raykov model predicts that allowing control decreases the subject's score in the knowledge test after the task (Movie B = -0.527, Diner B = -0.563). The coefficient was found to be significant in both the Movie study (p = 0.002) and the Diner study (p = 0.004), thus we reject H_3.

H_4: Control over DSS along with explanations leads to increased knowledge over time. In both studies, the Raykov change model predicts that allowing control along with showing explanations decreases the subject's score in the knowledge test after the task. Both predictors independently decrease the subject's score, however, their interaction increases their score. The solution by adding the "explanation" and "control" predictors and their interaction results in a net decrease of knowledge score (Movie solution = -0.276, Diner solution = -0.459). The coefficient was found to be significant in both the Movie study (p = 0.026) and the Diner study (p = 0.037), thus we reject H_4.

V. Discussion

In this research, we found that the data supported rejecting all of our initial hypotheses. However, we found a very interesting effect: our opaque, non-customizable systems were the best for participant learning. We discuss this below.

The knowledge complacency effect was observed in both studies, but only under specific conditions of explanation, control, and error. In the Movie study, there was an overall drop in knowledge score from 45.0% correct to 42.5% correct on average, regardless of experiment treatment (Figure 5). In the Diner study, there was an overall increase from 73.5% to 77.3% (all treatments, Figure 6). This is echoed in the original Diner's Dilemma study [14], where participant knowledge also increased as the game progressed. In both the Movie and Diner study, certain configurations of explanation, control, and error inhibited or affected learning. To better conceptualize the differences between treatments, we calculated the solutions to the multiple regression model, which is shown in Table II. These predictions imply that more user interaction with the agent (via control) leads to decreased knowledge. For instance, in the Movie study, our

TABLE I
FITTED RAYKOV MODEL COEFFICIENTS FOR MOVIE AND DINER STUDIES. THE BASELINE FACTOR ESTIMATES WERE PREDICTABLY NON-SIGNIFICANT (TREATMENT WAS INDEPENDENT OF USER KNOWLEDGE). BOLD INDICATES SIGNIFICANT PREDICTORS AT THE 95% CONFIDENCE LEVEL.

	Movie				**Diner**			
	B	Std. Error	Z-value	p (>|z|)	B	Std. Error	Z-value	p (>|z|)
Baseline Factor								
expl	-0.120	0.175	-0.688	0.492	0.099	0.190	0.519	0.603
cont	0.062	0.169	0.368	0.713	0.138	0.197	0.703	0.482
err	-0.141	0.171	-0.829	0.407	0.236	0.213	1.105	0.269
expl+cont	0.160	0.246	0.651	0.515	0.059	0.273	0.217	0.828
expl+err	0.167	0.248	0.672	0.502	0.066	0.296	0.222	0.824
cont+err	0.321	0.243	1.320	0.187	0.182	0.302	0.602	0.547
expl+cont+err	-0.489	0.349	-1.400	0.162	-0.473	0.424	-1.114	0.265
Change Factor								
expl	-0.298	0.175	-1.707	0.088	-0.464	0.190	-2.438	**0.015**
cont	-0.527	0.169	-3.128	**0.002**	-0.563	0.197	-2.858	**0.004**
err	-0.391	0.171	-2.293	**0.022**	-0.477	0.213	-2.234	**0.025**
expl+cont	0.549	0.246	2.231	**0.026**	0.568	0.273	2.081	**0.037**
expl+err	0.501	0.248	2.017	**0.044**	0.327	0.296	1.103	0.270
cont+err	0.571	0.243	2.349	**0.019**	0.533	0.302	1.761	0.078
expl+cont+err	-0.714	0.349	-2.043	**0.041**	-0.500	0.424	-1.178	0.239

TABLE II
SOLUTIONS TO THE CHANGE FACTOR MULTIPLE REGRESSION FOR EACH STUDY.

	Movie Study	**Diner Study**
Baseline	0.0 (at mean)	0.0 (at mean)
Explanation	-0.298	-0.464
Control	**-0.527**	**-0.563**
Error	-0.391	-0.477
Explanation + Control	-0.276	-0.459
Explanation + Error	-0.188	**-0.614**
Control + Error	-0.347	-0.507
Explanation + Control + Error	-0.309	**-0.576**

data showed that participants were much more likely to use the agent if control features were available. Increased interaction with the recommendation agent could have lead to a skewed perception of what was in the movie database – namely, a horror-movie aficionado may only see horror movies in the recommender, perhaps subtly convincing him that horror movies are the most highly rated genre (thus influencing the scores for knowledge questions 1 and 2 in Figure 5). Table II also supports the notion that over-trust may be a serious issue. For instance, in the Diner study, the error-prone version of the agent would exhibit a "flip-flop" behavior, where it changed its answer almost every round, making the sense-making process difficult. The explanations and control features that were provided only exacerbated the issue, as the estimates for those conditions are even lower.

The primary difference between the studies presented here and the original Diner study [14] was in the presentation of the agent: the original study framed the interface as a tool while the present studies framed the interface as an agent. Moreover, the control + explanation treatment (no error) was nearly identical to one of the treatments (UI Level 3). Upon comparison, we found that this treatment caused decreased learning in the present Diner study. In fact, the best treatment for learning in the Diner study was the no explanation, no control, no error treatment – a treatment which had a fairly static interface (the Dining Guru would "lock-in" to the best answer early on and then stay there). In contrast, the original Diner's Dilemma study saw equivalent levels of learning in all treatments. This suggests that it may be a combination of the agent framing, combined with features that make an agent appear competent (explanation, control) that trigger the knowledge complacency effect.

How can automated agents prevent knowledge bias? We note that of the two studies described here, two types of invocation strategies were attempted: always on (Movie) and on-demand (Diner), however, other types of invocation methods and explanation strategies have been studied in expert systems research [35]. While we initially believed prompting the user for input to get them involved in the process (Diner) or allowing the user to customize the agent's output (Movie) would mitigate any bias, it has become clear that more complex interaction strategies may be required. Providing recommendations adaptively, rather than making them available all the time, may be part of the solution, but further research as well as novel methods for interaction may be required.

Does knowledge complacency really matter? The effect sizes found in this study tended to be small, at most half of a standard deviation. Even if the effect sizes were large, does it really matter if operators are making correct decisions? We provide an argument to answer each of these two questions. First, small effect sizes can imply high cost depending on the domain. For instance, in movie recommendation, one user forming a misconception about a movie database might not be particularly devastating, however, if the user is a command and control operative and the domain is high risk, small policy changes can save hundreds of lives. Second, computational systems are not perfect and are (currently [36]) not legally liable for poor decisions, meaning for the foreseeable future humans will always be part of the decision loop. If agents make errors or fail electronically, humans will suddenly become responsible for any task they might have

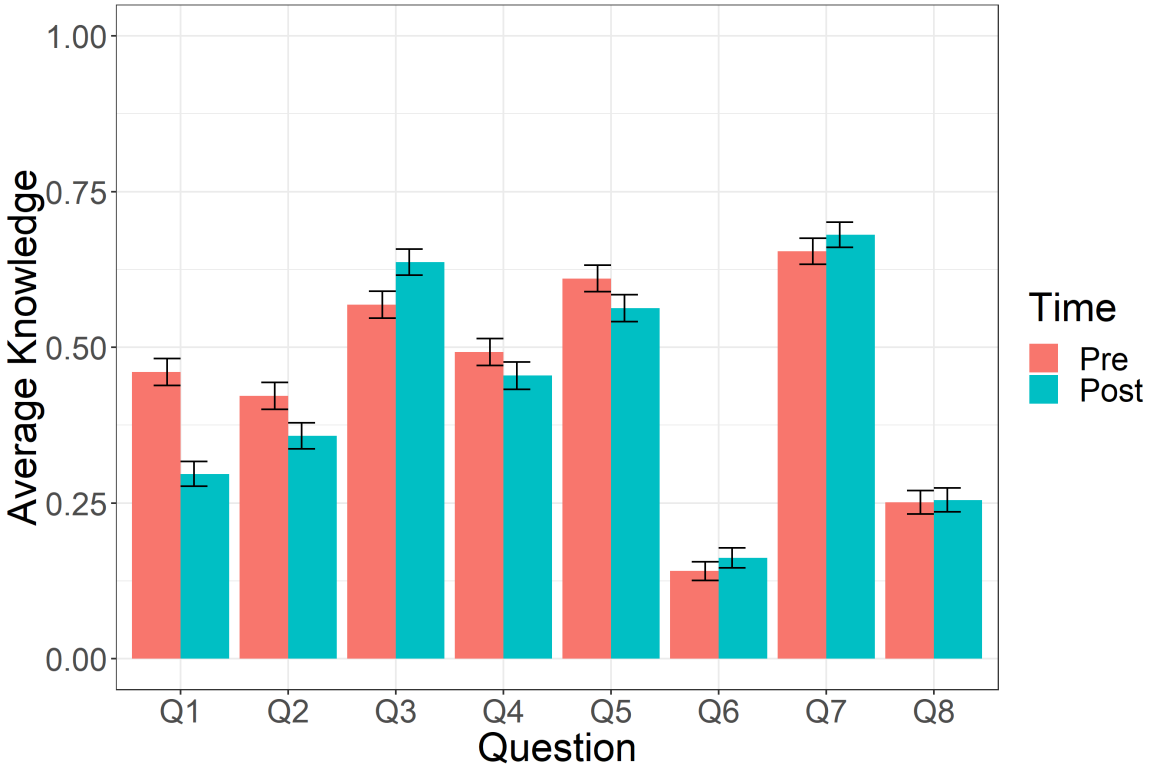

Fig. 5. Mean correct answers for each knowledge question in the Movie study. Error bars are 95% confidence intervals. Red bars indicates pre-test results, blue bars indicates post-test results.

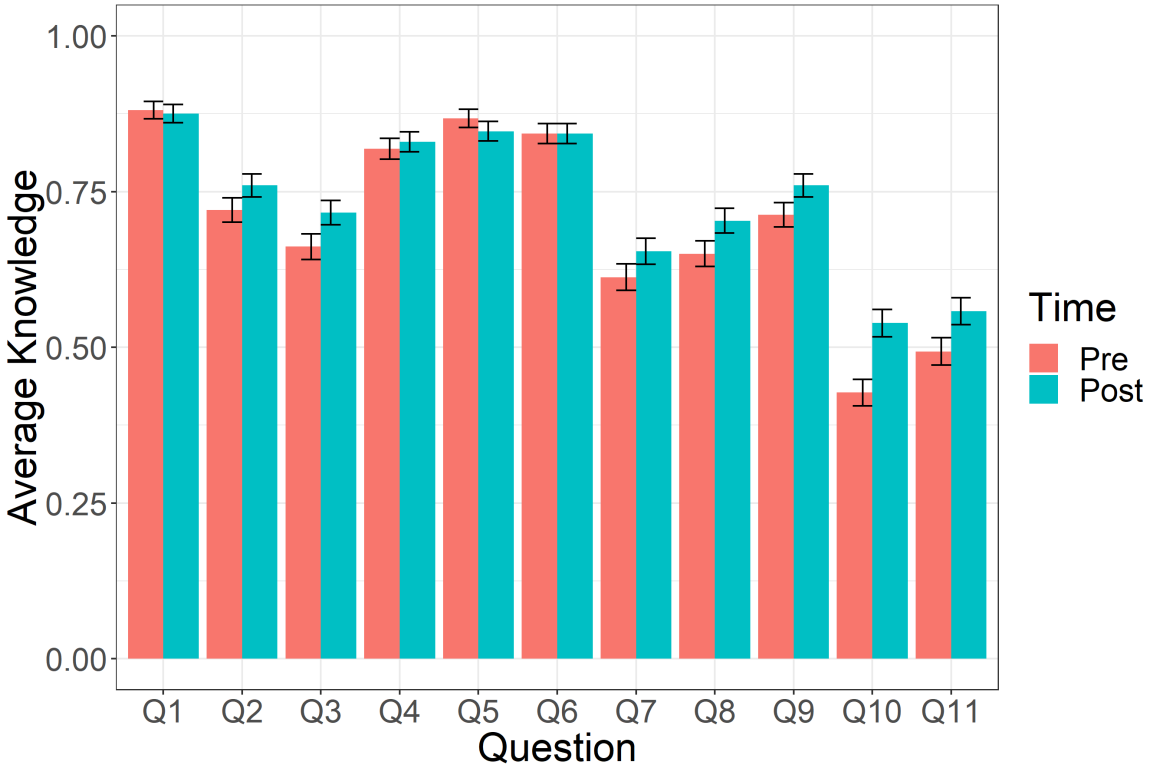

Fig. 6. Mean correct answers for each knowledge question in the Diner study. Error bars are 95% confidence intervals. Red bars indicates pre-test results, blue bars indicates post-test results.

been automating. The results from this study may make us think twice before blinding following the directions given by our GPS when travelling in unknown areas.

VI. Conclusion

Good decisions require users to internalize the output from information systems as knowledge. Complex decision aids that automatically summarize information can increase human capacity to make decisions. Automation complacency is recognized as a significant issue, however, the impact that these systems may have on human knowledge transfer and accumulation has not been thoroughly studied. We performed a comparative analysis between two studies that share a similar methodology and described how human knowledge can be negatively affected by the interface features of automated decision aids. This research has taken a small step towards a theory of how human-machine systems can become smarter and more effective.

References

[1] N. Tintarev and J. Masthoff, "A survey of explanations in recommender systems," in *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*. IEEE, 2007, pp. 801–810.

[2] V. Arnold, N. Clark, P. A. Collier, S. A. Leech, and S. G. Sutton, "The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions," *Mis Quarterly*, pp. 79–97, 2006.

[3] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa, "Inspectability and control in social recommenders," in *Proceedings of the sixth ACM conference on Recommender systems*. ACM, 2012, pp. 43–50.

[4] J. O'Donovan and B. Smyth, "Trust in recommender systems," in *Proceedings of the 10th International Conference on Intelligent User Interfaces*. ACM Press, 2005, pp. 167–174.

[5] R. Parasuraman, R. Molloy, and I. Singh, "Performance consequences of automation induced complacency," *International Journal of Aviation Psychology*, vol. 3, 02 1993.

[6] M. R. Endsley, *Designing for situation awareness: An approach to user-centered design*. CRC Press, 2011.

[7] T. B. Sheridan, *Human Supervisory Control*. John Wiley & Sons, Ltd, 2012, ch. 34, pp. 990–1015. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118131350.ch34

[8] R. K. Ahuja, K. Mehlhorn, J. Orlin, and R. E. Tarjan, "Faster algorithms for the shortest path problem," *Journal of the ACM (JACM)*, vol. 37, no. 2, pp. 213–223, 1990.

[9] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 43–52.

[10] J. E. Bahner, A.-D. Hper, and D. Manzey, "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience," *International Journal of Human-Computer Studies*, vol. 66, no. 9, pp. 688 – 699, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1071581908000724

[11] M. R. Endsley, "Automation and situation awareness," in *Automation and human performance*. Routledge, 2018, pp. 183–202.

[12] J. Schaffer, J. O'Donovan, and T. Höllerer, "Easy to please: Separating user experience from choice satisfaction," in *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*. ACM, 2018, pp. 177–185.

[13] E. Onal, J. Schaffer, J. O'Donovan, L. Marusich, M. S. Yu, C. Gonzalez, and T. Hollerer, "Decision-making in abstract trust games: A user interface perspective," in *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2014 IEEE International Inter-Disciplinary Conference on*. IEEE, 2014, pp. 21–27.

[14] J. Schaffer, J. ODonovan, L. Marusich, M. Yu, C. Gonzalez, and T. Höllerer, "A study of dynamic information display and decision-making in abstract trust games," *International Journal of Human-Computer Studies*, vol. 113, pp. 1–14, 2018.

[15] R. Parasuraman and D. H. Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 52, no. 3, pp. 381–410, jun 2010. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720810376055

[16] I. Singh, R. Molloy, and R. Parasuraman, "Automation induced "complacency": Development of the complacency-potential rating scale," *International Journal of Aviation Psychology - INT J AVIAT PSYCHOL*, vol. 3, pp. 111–122, 04 1993.

[17] K. Mosier and L. Skitka, *Human Decision Makers and Automated Decision Aids: Made for Each Other?* Erlbaum, 01 1996, vol. 40, pp. 201–220.

[18] J. Schaffer, P. Giridhar, D. Jones, T. Höllerer, T. Abdelzaher, and J. O'Donovan, "Getting the message?: A study of explanation interfaces for microblog data analysis," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 2015, pp. 345–356.

[19] M. Hertzum, H. H. Andersen, V. Andersen, and C. B. Hansen, "Trust in information sources: seeking information from people, documents, and virtual agents," *Interacting with computers*, vol. 14, no. 5, pp. 575–599, 2002.

[20] S. Y. Komiak and I. Benbasat, "The effects of personalization and familiarity on trust and adoption of recommendation agents," *MIS quarterly*, pp. 941–960, 2006.

[21] A. Cafaro, H. H. Vilhjálmsson, T. Bickmore, D. Heylen, K. R. Jóhannsdóttir, and G. S. Valgarsson, "First impressions: Users judgments of virtual agents personality and interpersonal attitude in first encounters," in *International conference on intelligent virtual agents*. Springer, 2012, pp. 67–80.

[22] S. Choi and R. E. Clark, "Cognitive and affective benefits of an animated pedagogical agent for learning english as a second language," *Journal of educational computing research*, vol. 34, no. 4, pp. 441–466, 2006.

[23] G. Veletsianos, "Cognitive and affective benefits of an animated pedagogical agent: Considering contextual relevance and aesthetics," *Journal of Educational Computing Research*, vol. 36, no. 4, pp. 373–377, 2007.

[24] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, D. Keim *et al.*, "Knowledge generation model for visual analytics," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, no. 12, pp. 1604–1613, 2014.

[25] P. Saraiya, C. North, and K. Duca, "An insight-based methodology for evaluating bioinformatics visualizations," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 11, no. 4, pp. 443–456, 2005.

[26] C. North, "Toward measuring visualization insight," *Computer Graphics and Applications, IEEE*, vol. 26, no. 3, pp. 6–9, 2006.

[27] V. Arnold, P. A. Collier, S. A. Leech, and S. G. Sutton, "Impact of intelligent decision aids on expert and novice decision-makers judgments," *Accounting & Finance*, vol. 44, no. 1, pp. 1–26, 2004.

[28] D. P. Ausubel and M. Youssef, "The effect of spaced repetition on meaningful retention," *The Journal of General Psychology*, vol. 73, no. 1, pp. 147–150, 1965.

[29] J. D. Karpicke and H. L. Roediger, "The critical importance of retrieval for learning," *science*, vol. 319, no. 5865, pp. 966–968, 2008.

[30] Internet Movie Database. https://www.imdb.com/.

[31] T. Raykov, "Structural models for studying correlates and predictors of change," *Australian Journal of Psychology*, vol. 44, no. 2, pp. 101–112, 1992.

[32] lavaan: latent variable analysis. http://lavaan.ugent.be/.

[33] The R Project for Statistical Computing. https://www.r-project.org/.

[34] R. Axelrod, *The Evolution of Cooperation: Revised Edition*. Basic Books, 2009. [Online]. Available: https://books.google.com/books?id=GxRo5hZtxkEC

[35] V. Arnold, N. Clark, P. A. Collier, S. A. Leech, and S. G. Sutton, "Explanation provision and use in an intelligent decision aid," *Intelligent Systems in Accounting, Finance and Management*, vol. 12, no. 1, pp. 5–27, 2004.

[36] Europe mulls treating robots legally as people... but with kill switches. https://www.theregister.co.uk/2017/01/13/eu_treat_robots_as_people/.