7 Quantitative Modeling of Dynamic Human-Agent Cognition

James Schaffer, James Humann, John O'Donovan, and Tobias Höllerer

CONTENTS

INTRODUCTION

Information systems have evolved to the point of being potential collaborators rather than manipulable tools. This has the potential to decrease human mental effort and increase the amount of data that can be incorporated into the human decision-making process. Intelligent agents can allow easy access to stored procedural knowledge and may alleviate the need to become an expert before taking action in a particular domain. Despite this, intelligent agents also face the danger of pushing the user out of the loop, such as pathfinding algorithms for automobile navigation (Ahuja, Mehlhorn, Orlin, & Tarjan, 1990) and collaborative filtering for movie recommendations (Breese, Heckerman, & Kadie, 1998), requiring only a yes-no confirmation but not revealing their underlying operations. Ideally, the complexity of these algorithms could be reduced to the level of common information tools such as winnowing interfaces, but this is not always possible. The conundrum of usefulness vs. simplicity was identified by Norman as early as 1986—he writes, "simple tools have problems because they can require too much skill from the user, intelligent tools can have problems if they fail to give any indication of how they operate and of what they are doing" (1986).

Designing interaction paradigms for intelligent agents remains an open problem (Gunning, 2017). The primary challenge is that most accurate algorithmic solutions for complex problems would require significant investment from a user to gain complete understanding. Even then, nonlinear decision boundaries utilized by an algorithm are difficult to visualize and explain, although there is progress on this front (Lakkaraju, Kamar, Caruana, & Leskovec, 2017; Ribeiro, Singh, & Guestrin, 2016a, 2016b). Another contributing factor is that algorithm technology continues to rapidly improve while models of human interaction and cognition during use of these systems lags behind. This might be because human-agent interaction (HAI) is a chaotic system (Gregersen & Sailer, 1993), making predictions of the convergence unlikely or impossible, even if ideal quantitative measurements could be taken. This

problem is further complicated by the potential of multi-agent systems, which stand to be even more difficult to model and harder to understand than single, "monolithic" systems. Despite these challenges, these problems can still be addressed, even if only uncertain or approximate solutions can be given (for example, predicting rain this afternoon with 51% or higher accuracy is much better than no prediction at all). This chapter defines and assesses the value of different cognitive and behavioral measurements in an attempt to explain variability in the human-agent system.

We propose profiling complex, automated algorithms using what we refer to as the explanation, control, and error (ECR) profile. We profile human users based on trust propensity, cognitive reflection, domain knowledge, and self-reported knowledge. We use the human and machine profile to investigate the human cognitive (trust, situation awareness, beliefs about an agent, perceptions of the agent, and cognitive load) and behavioral reactions to variations in these profiles. The factors investigated are then used in a statistical model to explain two types of human decision-making behaviors: adherence and decision outcomes. Specifically, we study how users interact with non-embodied, monolithic systems under two different task paradigms. We follow this with a discussion of how to extend the analysis to systems with multiple agents. Formally, we ask the following three questions about dynamic human-agent cognition:

(1) How do a person's cognitive traits affect usage of an intelligent agent and resulting decision outcomes?
(2) Which cognitive or system factors explain variability in decision making (interaction, adherence, success) in the HAI system?
(3) What is the relationship between correct beliefs about agents, their use, and trust?

In order to answer these questions, we generate a statistical map of all the factors mentioned through an exploratory factor analysis. We model the human-agent system by considering the agent's profile (explanation, control, error) as predictors of adherence and decisions while controlling for cognitive traits (domain knowledge, cognitive reflection, reported knowledge, and trust propensity). This results in the ability to predict the outcomes (decision making, adherence) in terms of the *starting point* of the HAI system. Moreover, we consider inter-task states and behaviors (perceptions, cognitive load, trust, situation awareness, interaction) as partial or full mediators of the starting point variables. The final measurement model is shown in Figure 7.1.

Two exploratory structural equation models (SEM) (Ullman & Bentler, 2003)—one from each study—were fit (by testing around 85 hypotheses). Controlling for multiplicity was done using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) using the exploratory value of $Q = 0.10$, which penalizes more for false positives than false negatives (in other words, we do not want to miss any potentially interesting effects that could be the basis for future studies). This quantitative map of these two studies will not only lead to better understanding of how humans react to intelligent agents, but also inform the design of future research in this area.

In this section we attempt to clearly define the semantic meaning of each factor studied. Then, we give a brief overview of the terminology used. Finally, we follow with a discussion of related work in each area.

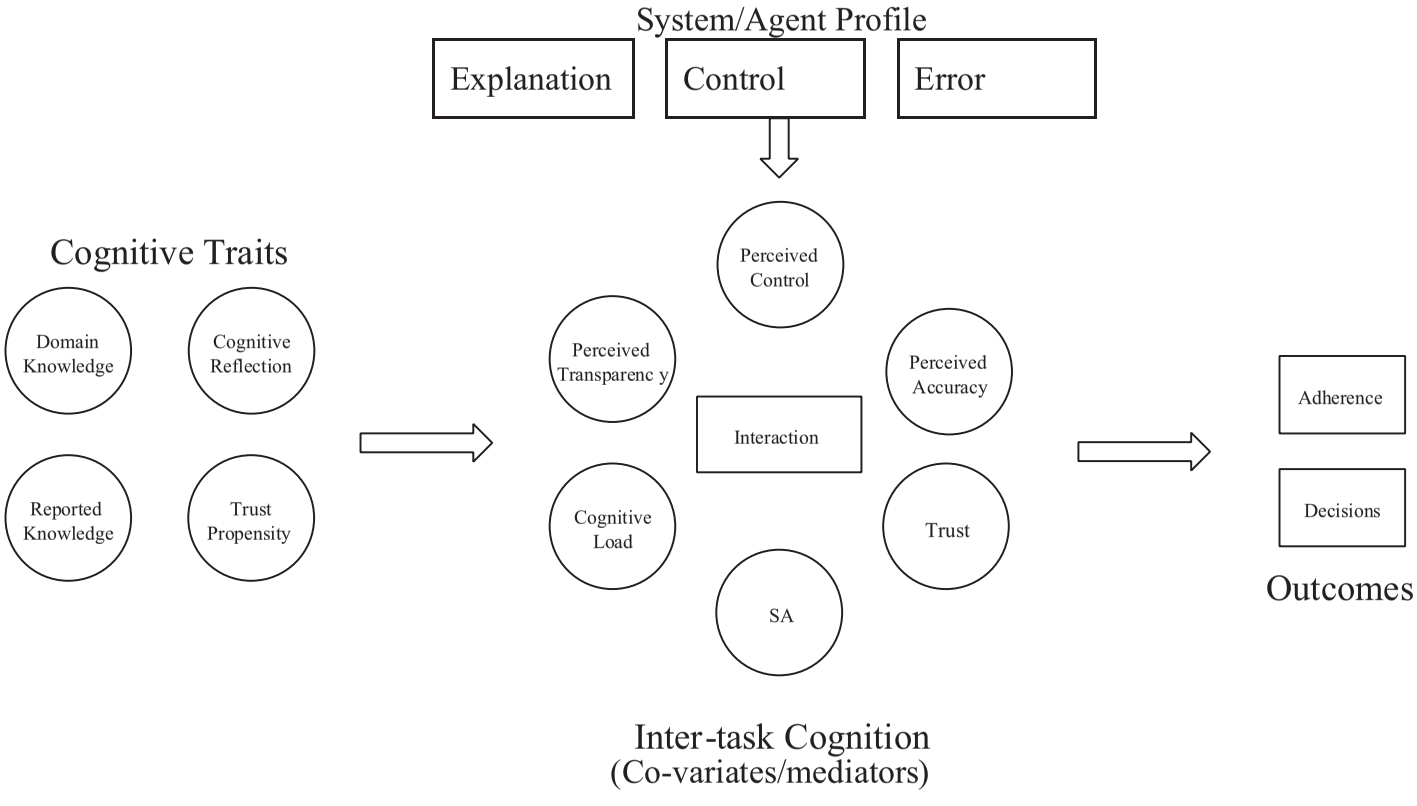

FIGURE 7.1 Measurement model used to quantify the human-agent system.

Example Application Domains

The context of this research is human interaction with intelligent agents when performing a task or making a decision. This interaction occurs in a wide and increasing variety of contexts as artificial agents' capabilities are bolstered by increasing access to data and computing power. We have alluded to familiar example use cases in route planning by a smartphone or GPS or movie recommendations based on inferred user preferences from online streaming providers such as Netflix. A more complex example is automated suggestions for disease diagnosis to supplement a doctor's expert analysis. Later in this chapter, we describe in detail a movie recommendation system and a restaurant ordering recommendation system to aid a user in maximizing the value he gets from a social meal. The common thread among all these examples is goal-oriented human interaction with an agent whose algorithms, reliability, and data store may not be fully transparent.

Terminology and Cognitive Factors

- Embodiment vs. artifact: We distinguish between two types of agents, those that are embodied visually or physically and those that are not (artifacts).
- Multi-agent systems: Multi-agent systems distinguish themselves from single-agent systems by maintaining separate internal states and knowledge of the environment.
- Subjective and objective task domains: An objective task domain has a criterion for success that can be measured and verified by a third party, such as the goal of removing body fat in the fitness domain. A subjective task domain attempts to model and satisfy the preferences of an individual person, such as the goal of providing an appropriate item to a customer.
- Explanation, control, and error: This is a simple method for profiling an agent. Explanation refers to the degree that operations are communicated to a user, control refers to the degree that agent behavior can be re-directed, and error refers to the probability that the agent's output does not solve the task domain.
- Situation awareness: Situation awareness is defined as the match between a person's mental model and the state of the environment. Situation awareness is defined either globally or with respect to a particular object in the environment.
- Trust, user experience, system perceptions: Trust is defined as a person's willingness to accept an agent's recommendations. User experience is the feeling (positive or negative) that a person has when interacting with an agent. System perceptions are defined as more nuanced forms of user experience, e.g., a person may have a good experience overall but may identify that an agent is bad at explaining itself.
- Cognitive load: A person's state of frustration while attempting to process task factors.

- Reported and true domain knowledge: Domain knowledge is defined as the total number of insights about a domain that a person has. Reported knowledge is their self-assessed knowledge level, which may not accurately reflect their true domain knowledge.

Embodiment vs. Artifact

Embodied agents are agents that are embedded within and control a particular physical system or visual entity, whether networked or un-networked (e.g., automated drones, Amazon's Alexa). Agents can also be a technological artifact of an algorithm (or collection thereof) that exhibits complex behavior but does not necessarily manifest in a physical or visual form (e.g., recommender systems on Amazon, Netflix, etc.). Virtual embodied agents (typically referred to as just virtual agents) are known to strongly influence the behavior of users (Hertzum, Andersen, Andersen, & Hansen, 2002) when compared to their nonembodied "artifact" counterparts (Komiak & Benbasat, 2006). This is because people react to virtual agents similarly to the way they react to real people. They form an opinion of the agent within the first 13 seconds of interaction and become more conscientious about behaviors (Cafaro et al., 2012). Despite this, trust relationships with non-embodied agents, especially recommender systems, continue to be studied (Benbasat & Wang, 2005; Knijnenburg, Bostandjiev, O'Donovan, & Kobsa, 2012a; O'Donovan & Smyth, 2005; Pu, Chen, & Hu, 2011), perhaps because virtual agents remain expensive and their performance is considered an open question (Choi & Clark, 2006; Veletsianos, 2007).

A physical embodied agent is tangible (e.g. drones, robots, and to some extent Amazon's Alexa), while intangible agents, such as the recommender systems that are embedded on modern e-commerce websites, reside in digital space. Human interaction with tangible agents can differ dramatically from intangible agents, even if the recommendations and overall system goal are the same. Interactions with physical agents can be colored by social cues, cultural norms, differing expectations relative to computers, and levels of acceptance of anthropomorphic form factors (Breazeal, 2004). It was shown in Podevijn et al., (2016) that humans have differing physiological effects when dealing with physical robot swarms than with virtual robot simulations. There is also evidence that users have a better subjective experience when performing a task aided by physical robots than by simulated robots (Wainer, Feil-Seifer, Shell, & Mataric, 2006). Humans have been shown to be more "polite" to physical robots than to virtual agents, being more likely to respond to greetings and afford the robot personal space while performing tasks (Bainbridge, Hart, Kim, & Scassellati, 2008). Perhaps most significantly, users may be more trusting of physical agents, as they were more likely to perform an unintuitive task when instructed by a robot than when instructed by a virtual agent (Bainbridge et al., 2008).

Monolothic vs. Multi-Agent Systems

Popular examples of monolithic systems include expert systems or recommender systems, which we expect many people to have come across on e-commerce sites such as Amazon or Netflix. There is a relative wealth of research on user behavior

and cognition with monolithic systems, which will be discussed in the following sections. Here, we focus on multi-agent systems, which is an emerging research area.

Multi-agent systems consist of multiple agents with separate internal states and knowledge of the environment. This knowledge shielding is critical to the definition of software multi-agent systems, as a group of software agents with perfect group knowledge and communication would be indistinguishable from a monolithic system (Panait & Luke, 2005). Systems can include multiple intangible agents, such as multi-disciplinary design aids (Ren, Yang, Bouchlaghem, & Anumba, 2011). Tangible multi-agent systems, such as groups of robots, are more easily recognizable as their practical restrictions on communication and distribution in space ensure that their states and knowledge cannot be uniform.

As autonomy becomes embedded in more objects of diverse form factors, humans are more likely to interact with multiple agents simultaneously. Increasing the number of agents in the system can place control and attention burdens on the human user, but there are also many benefits to distributed systems, such as adaptability and scalability (Humann, Khani, & Jin, 2016; Prokopenko, 2013; Requicha, 2013). In complex distributed systems, agents can reduce the cognitive load on humans by making decisions locally and at a lower hierarchical level. This is useful in many practical applications, such as power grid management or control of customizable manufacturing processes (Marik & McFarlane, 2005). This delegation to low-level agents allows the human to focus on decision making at a higher hierarchical level while tracking fewer details of the inner workings of the system. Other benefits include adaptability and re-configurability, as agents continuously adapt to one another and new agents that are introduced into the system. Agents can also support multi-disciplinary decision making, as each agent can represent specialized knowledge in a domain, relieving the user of the responsibility to be an expert in every relevant domain.

Multi-agent systems can be used as simulations, surrogates for the behavior of humans. This is especially helpful in design of systems that are meant to have many users simultaneously. This approach has been used to model organizational processes (Jin & Levitt, 1996), evacuation procedures (Mikhailov, 2011; Stuart et al., 2013), seating layout design (Humann & Madni, 2014), and many others. The results of simulation can predict how the systems will be used in practice, and allow for design changes to be made up front, rather than waiting for problems to arise in use.

In multi-agent intelligent systems, the intelligence can be distributed and devolved, so that the recommendations to the user are not coming from a single source. Recommendations may not always be meant for the user either; in a multi-agent system, agents could be making intelligent recommendations for use by other agents. For example, in Humann and Spero (2018), a surveillance task of classifying threats within a field was carried out by humans interacting with two different classes of unmanned aerial vehicles (UAVs). As a first pass, fast high-altitude vehicles would tag points of interest within the field. Their sensitivity to risk variables (e.g. heat, metal content, movement) could be increased to eliminate false negatives, but at the cost of increasing false positives. Thus the user must set the sensitivity at such a level that they can work efficiently, be confident that they are avoiding false negatives, and use further analysis to root out the false positives. At every step in this

process, there is an opportunity for agents to explain their recommendations to the user, but in large systems, this could quickly result in information overload.

When human control of each agent is infeasible, the agents must be made more autonomous or hierarchical control may be introduced. In one example of hierarchical control, "RoboLeader" is an autonomous virtual intermediary between a human operator and tangible robotic agents (Rosenfeld, Agmon, Maksimov, & Kraus, 2017; Snyder, Qu, Chen, & Barnes, 2010). The human interacts with RoboLeader by issuing high-level commands, and RoboLeader is responsible for controlling a team of robots that is searching for survivors in a simulated rescue mission. This hierarchy shields the complexity of multi-agent control from the human in most cases, while still allowing the human to take direct control of individual robots in special cases.

Explanation, Control, and Error

Research on virtual monolithic agents has led us to the theory that, at a fundamental level, all agents can be profiled by their levels of explanation, control, and error (ECR). Explanation level is the amount of output (and thus visual) bandwidth that is allocated for indications of operation. For instance, showing intermediate sorting steps would be an explanation of a sorting algorithm. Control level is the degree to which the system requires or allows input from the user. The ideas of control and automation are intrinsically linked. Increased automation necessarily leads to systems that more specifically target a particular task, reducing flexibility and reusability, but requiring less control. For instance, requiring the user to select the kernel of a support vector machine (as can be done in Weka; Holmes, Donkin, & Witten, 1994) decreases the level of automation and increases the cognitive demands on the user, but also increases the overall flexibility of the system. Explanation features are sometimes intentionally designed to accommodate control features, such as the selection of an alternate route in a GPS system. Automation can also be dynamic, turning on or off when the system detects it is in a critical state. Finally, all computational functions and algorithms solve a well-defined problem, but due to limitations in information or processing power, errors can occur. For instance, recommender algorithms attempt to predict user preferences in sets of items, but complete knowledge of a user's preferences can only be estimated from the user's item profile, which only partially defines their tastes. In other applications, processing may be under a time limit, which means systems must sometimes settle for approximate solutions.

Explanation and control from automated algorithms has been studied since at least 1975 (Shortliffe et al., 1975). This section presents work on explanation and control features in three research areas: recommender systems, expert systems, and scientific computing. We will also survey research where the accuracy of decision support systems was experimentally manipulated.

Explanation in Recommender Systems

Over the last 15 years, research has shown that explanation of a recommender system's reasoning can have a positive impact on trust and acceptance of recommendations. Recent keynote talks (Chi, 2015) and workshops (O'Donovan et al., 2015) have helped to highlight the importance of usability. Many recommender systems

function as *black boxes*, providing no transparency into the working of the recommendation process, nor offering any additional information beyond the recommendations themselves (Herlocker, Konstan, & Riedl, 2000). This may negatively affect user perceptions of recommendation systems and the trust that users place in predictions. To address this issue, static or interactive/conversational explanations can be given to improve the transparency and control of recommender systems (Tintarev, Kang, Höllerer, & O'Donovan, 2015).

Bilgic and Mooney (2005) furthered this work and explored explanation from the promotion vs. satisfaction perspective, finding that explanations can actually improve the user's impression of recommendation quality. Later work by Tintarev and Masthoff (2007) surveyed literature on recommender explanations and noted several pitfalls to the explanation process, notably including the problem of confounding variables. This remains a difficult challenge for most interactive recommender systems (Tintarev et al., 2014), where factors such as user cognitive ability, mood and other propensities, experience with the interface, specific interaction patterns, and generated recommendations can all impact on the user experience with the system. Sinha and Swearingen (2002) noted that users liked and felt more confident about recommendations they perceived as transparent. The importance of system transparency and explanation of recommendation algorithms has also been shown to increase the effectiveness of user adoption of recommendations by Knijnenburg et al. (2012a).

Explanation in Expert Systems

Work in knowledge-based or "expert systems" has illuminated the effects of exposing explanations from complex agents. Gregor and Benbasat (1999) provide an excellent summary of the theory of crafting explanations for intelligent systems. User studies which test the effects of explanation typically vary explanation level and quantify concepts such as adherence or knowledge transfer. Key findings show that explanations will be more useful when the user has a goal of learning or when the user lacks knowledge to contribute to problem solving. The impact of explanation on both novices and experts has also been extensively studied (Arnold, Clark, Collier, Leech, & Sutton, 2006): novices are much more likely to adhere to the recommender/expert system due to a lack of domain knowledge, and expert users require a strong "domain-oriented" argument before adhering to advice. Experts are also much more likely to request an explanation if an anomaly or contradiction is perceived. Most of these studies focus on decision-making domains (financial analysis, auditing problems) and were conducted before the explosion of data that is available to modern tools. When browsing or analyzing data that is too large to be analyzed by hand, decision makers have no choice but to utilize automated filtering techniques as part of their search strategy. This creates new questions about what might change in the dynamics between humans and automated algorithms.

Automation and Error

Intelligent assistants often vary in their degree of automation and effectiveness, but these have not garnered as much attention as the explanation issue. The pros and

cons of varying levels of automation have been studied in human-agent teaming (Chen & Barnes, 2014). 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. Koren & Bell, 2015), 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. (2015). In contrast, Yu et al. (2017) found that system errors do have a significant effect on trust and Harman et al. (2014) found that users trust inaccurate recommendations more than they should. 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.

Situation Awareness

The theory of situation awareness (SA) can answer some questions about human decision making in contexts where intelligent agents are present (Endsley, 1995b; Parasuraman, Sheridan, & Wickens, 2008). Maximal SA is a requirement for optimal decision making. If an analyst cannot understand what an intelligent agent is doing and an error is made, it could potentially result in catastrophic errors. For example, the Air France 447 crash[1] was caused by a combination of system error and lack of transparency. Measurement methodologies for SA have been established (Endsley, 1995a), although new SA question items must be devised for each new domain.

SA-Based Agent Transparency

The theory of SA has also been applied to the problem of agent transparency (Chen et al., 2014). Chen's theory is called SA-based agent transparency (SAT), which is based on Endsley's three levels of SA and other theories. Chen refers to Endsley's SA as "global" SA, while SAT is relevant only to transparency requirements relevant to understanding the intelligent agent's task parameters, logic, and predicted outcomes.

Incorporating all three levels of SA into SAT should help a user gain understanding of an agent's reasoning and operation and help the user make informed decisions about "intervention," or what we call here as the manipulation of a "control" parameter. Chen notes that automation reliability strongly influences a user's attitude toward automation which can have significant impacts on trust, and thus has an impact on the degree to which that automation is leveraged. Over-trusting automation leads to automation bias (Cummings, 2004) and under-trusting results in disuse of the automation. Chen notes that information visualization and the display of uncertainty are key factors in understanding automation and discussed this in more detail in Chen, Barnes, and Harper-Sciarini (2011a).

Trust, User Experience, and System Perceptions

The word "trust" has been used to describe a number of phenomena in many different domains and therefore it is carefully defined in this section. In this work, the word trust refers to the user's *perception* that he or she can blindly rely on the system. This view was

strongly influenced by the research of McKnight, which distinguishes the concepts of trust in technology and trust in other people (McKnight, Carter, Thatcher, & Clay, 2011; McKnight, Carter, & Clay, 2009), showing that people can discriminate between the two.

Trust propensity and its relationship to trust has been studied extensively in psychology, notably Colquitt et al. (2007) and Gill et al. (2005). Behavioral outcomes are affected by trust propensity when partially mediated by trust and *trustworthiness*, which is information about a trustee. The effects of trust propensity on behavioral outcomes disappears when information about the trustee becomes more reliable. Other studies in e-commerce have also found similar mediating effects between trust and trust propensity (Lee & Turban, 2001). Both trust and trust propensity need to be measured simultaneously to isolate system properties that instill trust from effects caused by highly trusting users.

Cognitive Load

The term "cognitive load" originates from education and learner theory (Sweller, 1994) and problem solving (Sweller, 1988) and is loosely defined as a "multidimensional construct representing the load that performing a particular task imposes on the learner's cognitive system." Information overload (Eppler & Mengis, 2004), a related concept, shares many of the same properties. Greater cognitive effort by users of systems leads to increased error when performing tasks. Paas et al. (2003) surveys numerous methods of measuring cognitive load during participant tasks, noting that cognitive load can be assessed by measuring mental load (portion of cognitive load that originates from task to subject relationship characteristics), mental effort (the actual effort exerted as demanded by task requirements), and performance. Participant self-reported rating scale techniques have been successful, as participants seem capable of accurately reporting their mental burden. Physiological techniques, such as the measurement of heart rate, brain activity, and pupil dilation, have also been successful. Finally, other kinds of performance measures can be applied, such as measuring the participant's effectiveness at managing a secondary task periodically while performing the primary task.

Cognitive Reflection

Work on attention and cognitive reflection (CRT) by Daniel Kahneman (1973) has been successful in discriminating between "fast" and "slow" thinking using a variety of questions that effectively trick the human processing system. Since then, CRT tests have been frequently used due to a correlation with human intelligence and decision making (Toplak, West, & Stanovich, 2011; Welsh, Burns, & Delfabbro, 2013). This work hypothesizes that CRT would be a strong predictor of a person's decision behavior when interacting with an agent.

Reported and True Domain Knowledge

Consequences of self-reported ability have been recently discovered in studies of cognitive psychology (Hoorens, 1993; Kruger, 1999). The Dunning-Kruger effect

predicts that low-ability individuals maintain an over-estimated 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 the user's actual ability in that area (Merritt, Smith, & Renzo, 2005). 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 (Kruger & Dunning, 1999).

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) (Davidson & Downing, 2000) 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 over-estimation of their own ability and their inability to assess the accuracy of the system.

Introduction to Task Paradigms

We considered two tasks to fit and validate the cognitive measurement model presented here. The first was a subjectively validated task—Movie Recommendation (MR)—and the second was an objectively validated game theoretic task—the Diner's Dilemma (DD).

In the MR task participants interacted with an interface dubbed "Movie Miner" to find a set of movies to watch in the future. This is a common setup in studies of recommendation, however, we improve upon these studies by including better modeling of decision satisfaction (Schaffer, O'Donovan, & Höllerer, 2018) and behavioral (rather than reported; Pu et al., 2011) adherence. Behavioral adherence modeling is only possible if the task is unrestricted and participants have the freedom to choose between alternative tools. Thus, participants were given two tools to work with and their behavior was not restricted. The methodology was thus very similar to typical online browsing sections, such as on Amazon, where a customer is browsing a product catalog and adding items to their "shopping cart." In summary, we kept the following three goals in mind: (1) to make the system as familiar to modern web users as possible, (2) to make the system as similar to currently deployed recommender systems as possible, and (3) to ensure that the study can be completed without forcing the users to accept recommendations from the system, so adherence can be measured. The use of novelty in any design aspect was minimized so that results would have more impact on current practice.

The second study, DD, was chosen 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 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 can show evolution in player strategies as they learn the other player's tendencies to defect or cooperate.

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 (Gonzalez, Ben-Asher, Martin, & Dutt, 2013; Martin, Juvina, Lebiere, & Gonzalez, 2011). In this experiment, the iterated three-player version was used, which limits the complexity of the game such that it is within the comprehension of human players, but is still sufficiently complex to warrant a computational aid. In this chapter, our DD gives the user a choice between ordering a hot dog or lobster when dining out with friends, under the agreement that the table's total bill will be split evenly among the diners. The *defect* strategy is to order the expensive and satisfying lobster, hoping that others will order hot dogs and subsidize the user's bill. The *cooperate* strategy is to order the inexpensive hot dog.

Participants in both studies were recruited on Amazon Mechanical Turk (AMT). AMT is a web service that gives tools to researchers who require large numbers of participants and are capable of collecting data for their experiment in an online setting. AMT has been studied extensively for validity; notably Buhrmester, Kwang, and Gosling (2011) found that the quality of data collected from AMT is comparable to what would be collected from laboratory experiments (Hauser & Schwarz, 2015). Furthermore, since clickstream data can be collected, satisficing—the act of rapidly "tab-clicking" through study questionnaires—is easy to detect.

MOVIE RECOMMENDATION METHODOLOGY

This section details the methodology used in the Movie Recommendation (MR) study.

Task and User Interface Design

The MR interface was closely modeled after modern movie "browsers" (such as IMDb or MovieLens) that typically have recommender functionality. On the left side of the interface, the system featured basic search, sort, and filter for the entire movie dataset. The right side of the interface provided a ranked list of recommendations derived from collaborative filtering, which interactively updated as rating data was provided.

The user interface provided the following functionality: mousing over a movie would pop up a panel that contained the movie poster, metadata information, and a plot synopsis of the movie (taken from IMDb); for any movie, users could click anywhere on the star bar to provide a rating for that movie, and they could click the green "Add to Watchlist" button to save the movie in their watchlist (CS was measured on their chosen movies at the end of the task). Clicking the title of any movie would take a user to the IMDb page where a trailer could be watched (this was also available during the CS feedback stage).

On the left (browser) side of this interface, users had three primary modes of interaction which were modeled after the most typical features found on movie browsing websites:

(1) Search: Typing a keyword or phrase into the keyword matching box at the top of the list returned all movies that matched the keyword. Matches were not personalized in any way (a simple text matching algorithm was used).
(2) Sort: Clicking a metadata parameter (e.g. Title, IMDb Rating, Release Date) at the top of the list re-sorted the movies according to that parameter. Users could also change the sort direction.
(3) Filter: Clicking "Add New Filter" at the top of the list brought up a small popup dialog that prompted the user for a min, max, or set coverage value of a metadata parameter. Users could add as many filters as they wanted and re-edit or delete them at any time.

The recommendation side operated identically to the browser side, except that the list was always sorted by the collaborative filtering prediction and the user could not override this behavior.

When explanations were provided, they appeared on mouse over and could not be hidden. First, each explanation stated: "Movie Miner matches you with other people who share your tastes to predict your rating." The rest of the explanation was generated by examining the three users in the database that were most similar to the user at the current point in time and taking the intersection of their rated movies with the user's profile. This identifies the movies that are *most responsible* for an item appearing at its respective location in the recommendation list.

The MovieLens 20M dataset was used for this experimental task. The MovieLens dataset has been widely studied in recommender systems research (Miller, Albert, Lam, Konstan, & Riedl, 2003; Jung, 2012; Harper & Konstan, 2016). Due to update speed limitations of collaborative filtering, the dataset was randomly sampled for 4 million ratings, rather than the full 20 million.

Agent Design

A traditional user-user collaborative filtering approach was chosen for the agent. Details for this can be found in Resnick et al. (1994). Collaborative filtering was chosen due to the fact that it is well understood in the recommender systems community and it achieves extremely high performance on dense datasets such as MovieLens (Koren & Bell, 2015). The results from this study should generalize reasonably well to other collaborative-filtering-based techniques, such as matrix factorization and neighborhood models. We made two minor modifications to the default algorithm based on test results from our benchmark dataset: Herlocker damping and rating normalization.[2]

ECR Manipulation

Two levels of control, two levels of explanation, and two levels of recommendation error were manipulated. All manipulations (three parameters, two values taken,

$2^3 = 8$ manipulations) were used as between-subjects treatments in this experiment. Text-based explanations were chosen due to their similarity to real-world systems such as Netflix and Amazon. To add user control, we chose to allow users to define filters on the list of recommendations. This approach is similar to real-world systems that are currently deployed on MovieLens, IMDb, and so on. To vary recommendation error, noise was added to the algorithm. This approach was validated by verifying that the random noise was reducing accuracy by performing a five-fold cross validation on our ratings data set. The error-free recommender achieved a mean absolute error of 0.144, while the noisy version did considerably worse at 0.181 (nearly a 26% difference).

Explanation Manipulation

- Opaque (*Explanation* = 0): The opaque recommender simply provided the recommendations without any explanation.
- Justification (*Explanation* = 1): The justification explained how ratings were calculated with the following blurb: "Movie Miner matches you with other people who share your tastes to predict your rating." This was followed by a list of the items in the user's profile that most affected the recommendation.

Control Manipulation

- Automatic (*Control* = 0): The recommender would update and re-sort its recommendations automatically. Participants could only affect the recommender's behavior by changing their user profile.
- Customizable (*Control* = 1): On top of the partial control features, users were allowed to define custom filters on recommender results to narrow the recommendations. Additionally, users could remove individual movies (indicating they were "not interested") from the recommendation list.

Error Manipulation

- Collaborative Filtering (*Error* = 0): Collaborative filtering: user-user similarity, Herlocker damping, and normalized across the 0.5–5 star rating scale.
- Collaborative Filtering with Noise (*Error* = 1): A vector of noise (of up to two stars difference) was calculated at session start and the vector was added in to the recommendation vector before normalization. From the participant's perspective, the list of recommendations thus appeared to be reordered as affected by this noise.

Procedure

Participants made their way through four phases: the pre-study, the ratings phase, the watchlist phase, and the post-study (Figure 7.4, top). The pre-study and post-study were designed using Qualtrics.[3] In the "ratings" phase, participants accessed Movie Miner and were shown only the blue "Movie Database" list and the ratings box. We asked participants to find and rate *at least* ten movies that they believed would best represent their tastes, but many participants rated more than the minimum. In

the "watchlist phase," participants were shown the brown "Recommended for You" list and the watchlist box. Instructions appeared in a pop-up window and were also shown at the top of the screen when the pop-up was closed. Participants were told to freely use whichever tool they preferred to find some new movies to watch. They could add movies to their watchlist with the green button that appeared on each individual movie (regardless of the list that it appeared in). We asked them not to add any movies that they had already seen, required them to add at least five movies (limited to seven maximum), and required them to spend at least 12 minutes interacting with the interface. A 12-minute session in which five to seven items are selected was deemed sufficient time to select quality items, given that people only browse Netflix for 60 to 90 seconds to find a single item before giving up (Gomez-Uribe & Hunt, 2016).

Accommodating Subjective Decision Making

Quantifying decision satisfaction is problematic because it can be influenced by user experience, mood, health, and so on. To improve modeling of subjective decision satisfaction, we used the satisfaction baseline approach (Schaffer et al., 2018). Baseline satisfaction was measured shortly after the pre-study by getting participant feedback on movies that were chosen from the database at random.

Ten random movies were shown, one at a time, and the responses (question items, bs1–4, given in Table 7.4) were averaged together. Satisfaction with selected items was measured after the "watchlist" phase. For this, the recommender interface was removed and the questions items were shown for each item chosen by the participant. Note that the question items are phrased in terms of the recommendations (Table 7.4, ds1–4), not the interface. This is to help the participant distinguish between the browsing tools and the features of the recommender system. By modeling changes between baseline satisfaction and satisfaction with selected items, it is possible to quantify the *change* in satisfaction, which is what the user interface would actually influence.

DINER'S DILEMMA METHODOLOGY

This section details the methodology used in the Diner's Dilemma (DD) study. A screenshot of the interface used is shown in Figure 7.2.

TABLE 7.1
Diner's Dilemma Choice Payoff Matrix

Player Chooses:			
Hot Dog	**Lobster**		
Two co-diners	20.00	24.00	Cooperate
One co-diner	12.00	17.14	Cooperates
	8.57	13.33	Neither cooperates

Task and User Interface Design

In the Diner's Dilemma, several diners eat out at a restaurant 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 (hot dog) 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 hot dog 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. Hot dog/lobster cost and values for the game are shown in Figure 7.2, under each respective item, resulting in the payoff matrix that is shown in Table 7.1.

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 Choi, de Melo, Khooshabeh, Woo, & Gratch, 2015). 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 hot dog order with a lobster order) or forgive (respond to a lobster order with a hot dog 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 hot dog 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 Table 7.2.

Participants played the game through the interface shown in Figure 7.2. 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 7.3).

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

TABLE 7.2
Performance of the Dining Guru (DG) across All Games Compared to Participants

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	HOT DOG	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

Note: 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.

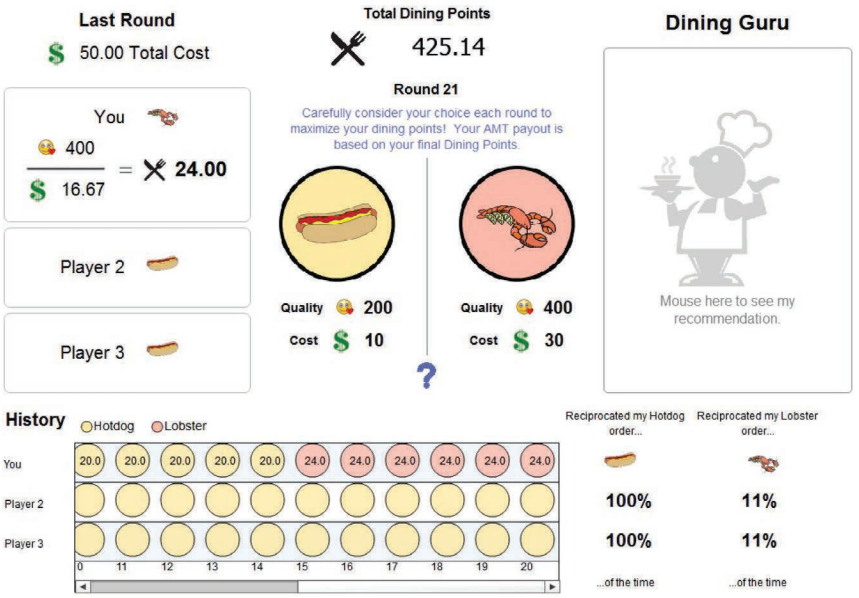

FIGURE 7.2 The user interface for the game.

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

Agent Design

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

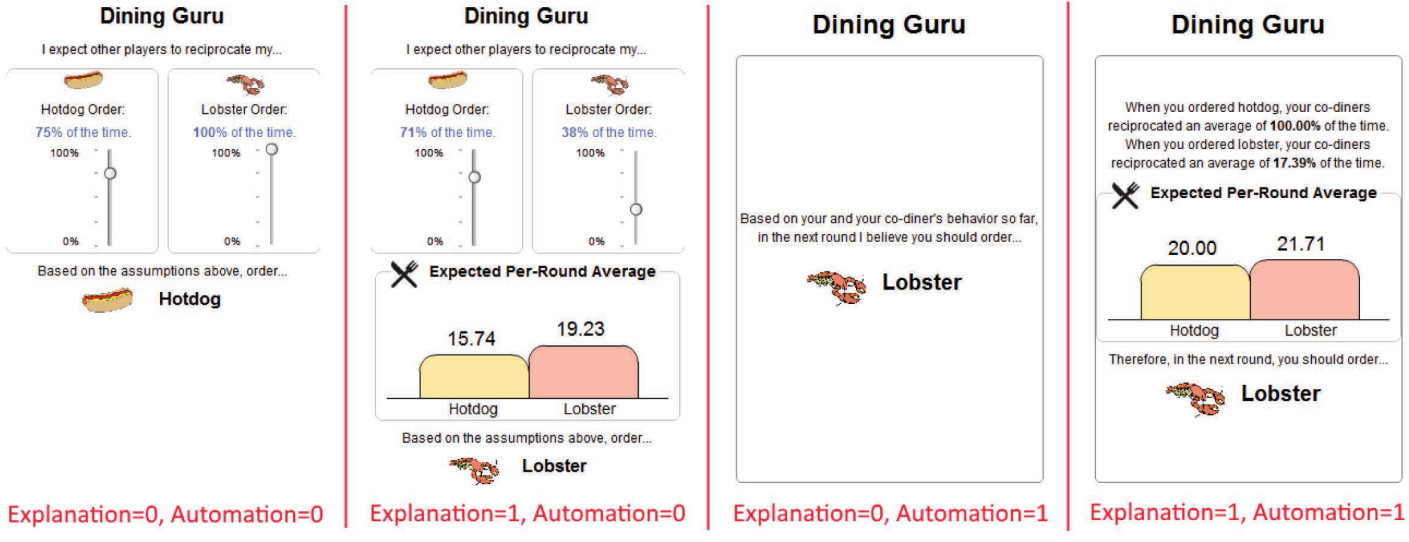

FIGURE 7.3 Variations of the Dining Guru, based on automation and explanation treatments.

about which item (hot dog 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 hot dog 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 hot dog and lobster reciprocity rates (see Figure 7.3).

ECR Manipulation

Two levels of control (*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 7.3 for a visual). All manipulations (three parameters, $3 \times 2^2 = 12$ 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. 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, a text blurb appeared explaining the connection between co-diner reciprocity rates and the expected per-round average. Reciprocity rates were provided by participants in the non-automated version, so the explanatory version only required the addition of the expected per-round averages. A bar graph was used to represent the averages so that participant attention would be drawn to the explanation.

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 mostly 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 hot dog reciprocity rate was 65%, the Dining Guru would use a value anywhere between 40% and 90%. 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.

Explanation Manipulation

- Opaque (*Explanation* = 0): The opaque recommender simply provided the recommendations without any explanation.

- Justification (*Explanation* = 1): The justification explained the relationship between co-diner reciprocity rates and optimal choices. This was communicated visually with a bar graph, showing the expected points per round for each choice based on historical co-diner data.

Control Manipulation

- Automatic (*Control* = 0): The recommender would update its recommendations automatically. Participants would have to mouse over the Dining Guru each round to check the most up-to-date recommendation.
- Customizable (*Control* = 1): The recommender required the user to provide estimated reciprocity rates for each co-diner. The estimates were provided by moving two sliders, which 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.

Error Manipulation

- Maximum Likelihood Estimation (*Error* = 0): Dining Guru produced recommendations that would be unbeatable, which if followed would result in the maximum number of optimal moves.
- Maximum Likelihood Estimation with Weak Noise (*Error* = 0.5): Randomly adjusts estimates for reciprocity rates up and down 25%, resulting in occasionally inaccurate recommendations.
- Maximum Likelihood Estimation with Noise (*Error* = 1.0): Randomly adjusts estimates for reciprocity rates up and down 50%, resulting in frequently inaccurate recommendations.

Procedure

An overview of the procedure for the DD study is given in Figure 7.4 (bottom). Before playing the game, participants were introduced to game concepts and the Dining Guru by playing practice rounds (training phase). 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 taking its advice was their choice. Participants played three games of Diner's Dilemma against three configurations of simulated co-diners with varying behavior characteristics.

TASK COMPARISON

This section details the differences in measurements and tasks between the DD and MR studies.

Item Operationalization

A comparison of the differing procedures is shown in Figure 7.4. This figure indicates where measurements were taken for each study, which was largely the same. The most critical difference is that the DD study contained a training phase, so the domain knowledge test measured retention rather than the participant's stored knowledge.

In the MR study, quantity (and type) of interaction with each tool was measured, constituting recommender and browser interaction. Adherence was measured as the percentage of items in each participant's watchlist that originated from the recommender side of the interface. Decision satisfaction was modeled as a two-wave, multi-item factor, so was modeled via confirmatory factor analysis (CFA). In the DD study, the quantity of interactions with the Dining Guru constituted recommender interaction. Absence of interaction with the Dining Guru was treated equivalent to browser interaction in the MR study. Adherence occurred for each round where the user choice matched the last recommendation given by the Dining Guru. The final adherence measurement was scaled between 0 and 1, where 0 indicates no adherence and 1 indicates complete adherence. Some users never accessed the Dining Guru, which caused their adherence score to become 0. Decision optimality was quantified as the total percentage of rounds where optimal decisions were made. An optimality score of 1 indicates the player ordered 100% lobster in games 1 and 3, and 100% hot dog in game 2.

We used a SAGAT-style freeze (Endsley, 2000) to assess situation-awareness-based agent transparency (SAT). For MR, this was done eight minutes into the watchlist phase of the task, whereas for DD it was done partway through game 2. Awareness of game factors (SAG) was also taken in the DD study. The SAG questionnaire contained five 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 hot dog (2, 3) and lobster (4, 5) reciprocity rates for each co-diner. The game interface was not available at this time. The SAG score was calculated by first summing up the errors from each of the five estimation questions and then inverting the scores based on the participant with the highest error, such that higher SAG scores are better.

CFA was used to eliminate measurement error when possible. Factor fit was improved iteratively by removing items until Cronbach's alpha was maximized, resulting in the list of items shown in Table 7.4. Internal reliability fit metrics for each factor are shown in Table 7.6. Domain knowledge, SAT, and SAG were expected to be multidimensional at the outset, and thus were parceled instead of factored—the question items used for the parcels is shown in Table 7.5. For the parceling, question items were summed and the loading of the factor on the parcel was set to 1. The variance of the parcel was freed to maximize fit.

In this study, we originally intended to model perceived transparency, control, effectiveness, and trust as separate factors. However, during the confirmatory factor analysis, inter-item correlations indicated we only had a single factor for MR and two factors for DD: perceived control and trust. Moreover, the perceived control factor complicated pathways in the model and was not strongly predictive of each outcome.

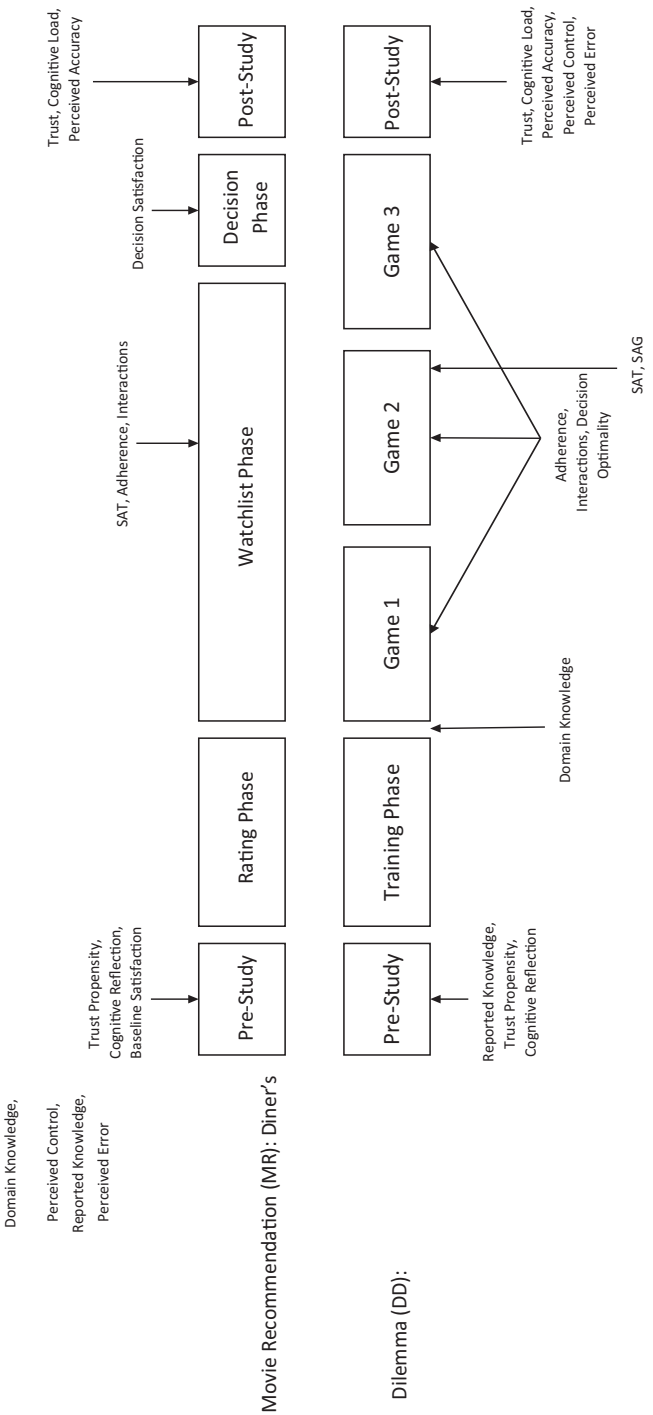

FIGURE 7.4 Overview of procedure and measurement timings for each study.

Thus, we collapsed these factors into a single factor, which we referred to as user experience with the agent (UXA). Doing so not only increased factor fit, but model fit as well. Once this was done, all factors in Table 7.6 achieved discriminant validity using the Campbell and Fiske test (Campbell & Fiske, 1959).

Task Differences

The main difference between the MR and DD studies lies in the nature of the task and the criterion for decision success. Both task spaces have been studied extensively. The DD task is a variant of the iterated Prisoner's Dilemma, whose applicability to real-world situations has been well established[4] (Stephens, McLinn, & Stevens, 2002; Ainslie, 2001; Varian, Bergstrom, & West, 1996). Decision success for each study was based on different parameters, with success in the MR study being subjective and success in the DD study being objective. It should also be noted that the treatment manipulations for explanation and control were minimal. This was done due to an understanding that decision makers are sensitive to the environment in which decisions are made (Payne, Bettman, & Johnson, 1993) and also to increase the relevance of the results (it is easier to implement a text-based explanation that a visual one). Differences between effects in the studies can thus be attributed to differences in the task parameters and decision criterion (Table 7.3), while similarities in effects thus have strong support for their generalization.

TABLE 7.3
Comparison of Task Parameters, Decision Success Criteria, and Treatment Differences between the Movie Recommendation (MR) and Diner's Dilemma (DD) Studies

Study	MR	DD
Decision task	Catalog browsing	Binary choice
Number of decision iterations	5–7	173
Agent support	Collaborative filtering	Maximum likelihood estimation
Alternative	Winnowing interface	History visualization
Embodiment	Artifact	Picture
Decision Criteria	Subjective	Objective
Domain	Movie metadata	Game rules
Explanation manipulation	Text-based explanation of agent's calculation	Text-based explanation of agent's calculation
Control manipulation	Optional metadata filters to be applied on ranked recommendation list	Requires specification of input parameters but allows exploration of metadata space
Error manipulation	Noise added to recommendation score, changing top recommendations	Noise added to expected values of binary choice, changing per-round recommendations

RESULTS

We collected more than 1,055 samples of participant data using Amazon Mechanical Turk: 526 samples for MR and 529 samples for DD. Participant data was checked carefully for satisficing and these records were removed (approximately ten per study), resulting in the 1,055 complete records. Participants were paid $3.00 for the DD and $1.50 for MR. In either case, participants spent between 25 and 60 minutes completing the study. Participants were between 18 and 71 years of age and were 50% male, however, DD attracted more male participants (54%) while MR attracted more female (55%).

Means and variances of non-factor measurements are given in Table 7.7 (factors are not listed here—all factors are modeled to have a mean of 0 and standard deviation of 1). Scores on tests were normalized between 0 and 1. Note that in the DD study, absence of interaction with the agent was considered interaction with the alternative. Decision optimality was modeled as two-wave decision satisfaction in the MR study, meaning that it was modeled as a factor and thus does not appear in the table.

Fit SEM Models

Data from each study was fit using an exploratory SEM, with the exception that decision satisfaction from the MR study was analyzed separately in a Raykov change model (Raykov, 1992) (Table 7.8). This is because baseline satisfaction needs to be taken into account when evaluating subjective satisfaction (Schaffer et al., 2018). A visual comparison of the results from both studies is shown in Figure 7.5, along with fit statistics and regression statistics. Due to being non-normal, treatment variables take the value of 0 or 1 and coefficients reported in the figure are B values (effect sizes in the units of the original measurement), which predict a change in standard deviation of the regressand when the treatment is switched on. All dependent and latent variables were standardized to have a mean of 0 and variance 1 and coefficients reported are β values, which predict a change in standard deviation of the regressand with a standard deviation change in the regressor. Both models were built using R 3.0.3, lavaan 0.5–17.

Multiplicity control was enforced in our chosen SEM using the Benjamini-Hochberg procedure with $Q = 0.10$ (Benjamini & Hochberg, 1995), which is recommended for exploratory SEM analysis (Cribbie, 2007). This procedure indicates how many of the tested relationships in the all-factor SEM are expected to be false positives. The MR and DD tasks had 86 and 87 hypotheses, respectively. These hypothesis numbers are derived from the exploratory way in which the SEMs were built, that is, specifying some factors/variables as downstream from others and testing the presence of significant predictive or causal relationships. Effects that failed the false discovery rate test were trimmed from the models—these were typically regressions on target variables whose regressor was already significantly correlated with another variable that predicted on the regressand. For example, reported and domain knowledge had a significant negative correlation in both studies. When controlling for one

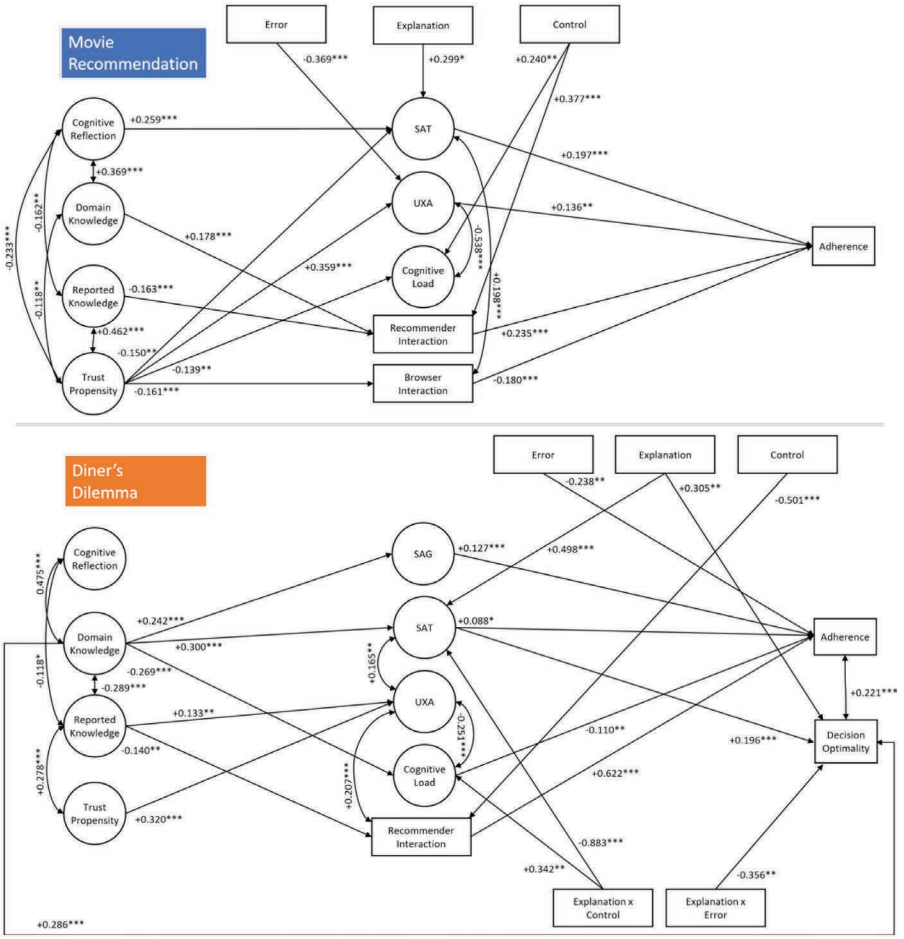

FIGURE 7.5 A comparison of the two fitted SEMs from each study. Movie Recommendation model fit: $N = 526$ with 77 free parameters = approximately 6.5 participants per free parameter, $RMSEA = 0.054$ (CI: [0.050, 0.057]), $TLI = 0.919$, $CFI = 0.926$ over null baseline model, $\chi^2(512) = 1285.408$. Diner's Dilemma model fit: $N = 529$ with 72 free parameters = approximately 7 participants per free parameter, $RMSEA = 0.030$ (CI: [0.025, 0.035]), $TLI = 0.969$, $CFI = 0.973$ over null baseline model, $\chi^2(378) = 557.889$.

or the other, effects on regressands (e.g., SAT) become less significant and fail the false discovery rate threshold.

DISCUSSION

This section first discusses the statistical effects observed in each study in detail. Then, we compare the results from two studies. Finally, we discuss the implications of these effects in the context of multi-agent and physical systems and highlight future research challenges.

Movie Recommendation: Statistical Effects

In the MR data, we found that splitting user experience into different subjective user-experience factors (similar to the ResQue framework (Pu et al., 2011) and the model in Knijnenburg et al. (2012a)) decreased model fit, despite each sub-factor (perceived transparency, perceived control, perceived quality, and trust) having items with acceptable fit but high inter-correlation (about 0.95), implying poor discriminant validity. Generally, high correlations among factors are undesirable due to the decreased questionnaire item-to-information ratio. For instance, in this study, a three-item scale for "trust" would have captured nearly the same signal as the 12-item SSA model that was used. This may have occurred because participants had a unidimensional perception of the recommender (i.e. "I like this" or "I don't like this"), which was a surprising finding. We considered it important to compare our results with Knijnenburg et al. and the ResQue framework. The Knijnenburg data was available[5] and we examined the covariances of perceived quality, satisfaction, control, and understandability. The scales in Knijnenburg's study were slightly better in terms of discriminatory power: about a 0.7 Pearson correlation between perceived overall system satisfaction, quality, and control, but this correlation level is still quite high. The transparency subconstruct, "understandability," is much more discriminative (0.34), perhaps due to the user-centric phrasings used. Unfortunately, discriminant validity between factors in the ResQue framework were not reported. For interested readers, a detailed discussion of user experience modeling along this vein is discussed in Schaffer et al. (2018). In light of this analysis, we encourage other researchers to consider the inter-factor correlations and discriminant validity of their chosen factors.

As evidenced by the profiling traits of CRT, domain knowledge, reported knowledge, and trust propensity, users of intelligent agents can be broken into two groups of high and low task ability. Higher-ability users are more likely to understand the recommender but less likely to form positive perceptions of it, while lower-ability users reported being more trusting and over-estimating their task knowledge, subsequently interacting and adhering to the system to a lesser degree and having worse decision outcomes overall.

Despite the recommender system community's emphasis on user experience, we found that user experience with the agent (UXA) was perhaps the most trivial factor in predicting adherence. This is evidenced in Figure 7.5, where it can be observed that SAT, recommender interaction, and browser interaction all are better predictors of adherence than UXA. The exception to this was cognitive load, which does not correlate with adherence in the final fitted model. However, cognitive load and user experience were strongly negatively correlated, so any alternative model using cognitive load as a predictor of adherence instead of UXA is valid. This result reinforces the idea that cognitive load and user experience have an inverse relationship (see Jung (2012)).

Curiously, UXA was a negative predictor of decision satisfaction, as was adherence to recommendations. This again contradicts results from other studies of recommendation (Knijnenburg, Willemsen, Gantner, Soncu, & Newell, 2012; Pu et al., 2011), however, this study has the advantage of modeling change in satisfaction over the baseline (Schaffer et al., 2018)—Δ Decision Satisfaction, rather than just satisfaction at one point in the task. This modeling is more accurate because it accounts for

TABLE 7.4
Factors Fit from Participant Responses to Subjective Questions

Code	MR Item	DD Item
re1	I am an expert on movies.	I am familiar with abstract trust games.
re2	I am a film enthusiast.	I am familiar with the Diner's Dilemma.
re3	I closely follow the directors that I like.	I am familiar with the public goods game.
crt1	If it takes five machines five minutes to make five widgets . . .	(same as MR)
crt2	A bat and ball together cost $1.10, and the bat costs $1.00 more than the ball . . .	(same as MR)
crt3	In a pond there is a patch of lily pads that doubles in size every day . . .	(same as MR)
tp1	I think I will trust the movie recommendations given in this task.	I think I would trust an AI adviser if one were available.
tp2	I think I will be satisfied with the movie recommendations given in this task.	I think I would be satisfied if I adhered to advice from an AI adviser.
tp3	I think the movie recommendations in this task will be accurate.	I think AI advisers give accurate information.
pt1	How understandable were the recommendations?	The Dining Guru's recommendations were understandable.
pt2	Movie Miner succeeded at justifying its recommendations.	I did not understand the Dining Guru.
pt3	The recommendations seemed to be completely random.	The Dining Guru's recommendations were groundless.
pa1	I preferred these recommendations over past recommendations.	
pa2	How accurate do you think the recommendations were?	The Dining Guru was accurate.
pa3	How satisfied were you with the recommendations?	The Dining Guru's recommendations were satisfactory.
pa4	To what degree did the recommendations help you find movies for your watchlist?	The Dining Guru's recommendations helped me to maximize points.
pc1	How much control do you feel you had over which movies were recommended?	I had control over the Dining Guru.
pc2	To what degree do you think you positively improved recommendations?	I could affect what the Dining Guru recommended.
pc3	I could get Movie Miner to show the recommendations I wanted.	I had no control over the Dining Guru.
t1	I trust the recommendations.	I trusted the Dining Guru.

TABLE 7.4 (Continued)
Factors Fit from Participant Responses to Subjective Questions

Code	MR Item	DD Item
t2	I feel like I could rely on Movie Miner's recommendations in the future.	I could rely on the Dining Guru.
t3	I would advise a friend to use the recommender.	I would advise a friend to take advice from the Dining Guru if they played the game.
cl1	There was too much information on the screen.	It was hard to keep track of all of the information needed to play the game.
cl2	I got lost when performing the task.	I got lost while playing the game.
cl3	Interacting with Movie Miner was frustrating.	I got frustrated during the game.
cl4	I felt overwhelmed when using Movie Miner.	
ds1	How excited are you to watch <movie>?	
ds2	How satisfied were you with your choice in <movie>?	
ds3	How much do you think you will enjoy <movie>?	
ds4	What rating do you think you will end up giving to <movie>?	
bs1	How excited would you be to watch <movie>?	
bs2	Would you be satisfied with choosing <movie>?	
bs3	How much do you think you would enjoy <movie>?	
bs4	What rating do you think you would end up giving to <movie>?	

Note: Items that were removed due to poor fit are not shown. All items achieved good fit, except for perceived transparency in the DD task, which was borderline.

TABLE 7.5
Question Items Used for Parceled Factors in Both Studies. Sum of Correct Responses Were Used to Calculate the Parcel

Code	MR Item	DD Item
dom1	Online, which genre has the highest current average audience rating?	In a one-round Diner's Dilemma game (only one restaurant visit), you get the least amount of dining points when . . . (four options)
dom2	Online, which of these genres tends to be the most common among the movies with the highest average audience rating?	In a one-round Diner's Dilemma game (only one restaurant visit), you get the most amount of dining points when . . . (four options)

(*Continued*)

TABLE 7.5 (Continued)
Question Items Used for Parceled Factors in Both Studies. Sum of Correct Responses Were Used to Calculate the Parcel

Code	MR Item	DD Item
dom3	Online, which of these genres has the highest current popularity?	Suppose you know for sure that your co-diners reciprocate your hot dog 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)
dom4	Generally, which of these genres has the most titles released, for all time periods?	Suppose you know for sure that your co-diners reciprocate your hot dog 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)
dom5	Online, which of these decades has the highest current average audience rating?	Suppose you know for sure that your co-diners reciprocate your hot dog 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)
dom6	How many movies have an average audience rating great than 9/10?	How much does a hot dog cost? (slider response)
dom7	Popular movies tend to have an average rating that is lower/ average/higher.	How much does a lobster cost? (slider response)
dom8	Movies with an average rating of 9/10 or higher tend to have fewer/ average/more votes.	What is the quality of a hot dog? (slider response)
dom9		What is the quality of a lobster? (slider response)
dom10		Which situation gets you more points? (two options)
dom11		Which situation gets you more points? (two options)
sat1	What is the recommender trying to predict?	The Dining Guru updates automatically every round. (T/F)
sat2	Are the recommendations I see just for me?	When the Dining Guru is updated, it predicts the choice I should make in the next round. (T/F)
sat3	What are the recommendations affected by?	When the Dining Guru is updated, it predicts the choice I should make in all remaining rounds. (T/F)
sat4	What are the recommendations based on?	When does the Dining Guru recommend hot dog?
sat5	When does the recommender update?	How does the accuracy of the Dining Guru change as the game progresses?
sat6	What happens if I delete all drama movies from my ratings?	Generally, I can maximize the dining points I get per round by ordering a mix of hot dog and lobster, regardless of what the Dining Guru recommends. (T/F)

TABLE 7.5 (Continued)
Question Items Used for Parceled Factors in Both Studies. Sum of Correct Responses Were Used to Calculate the Parcel

Code	MR Item	DD Item
sat7	What if I were to highly rate movies in the sci-fi genre?	Generally, I can maximize the dining points I get per round by only ordering what the Dining Guru recommends. (T/F)
sat8	What happens if I rate more movies according to my tastes?	
sat9	What happens if I remove accurate ratings?	
sag1		What is your current cooperation rate? (slider 0–100%)
sag2		What is Player 2's hot dog reciprocity rate? (slider 0–100%)
sag3		What is Player 2's lobster reciprocity rate? (slider 0–100%)
sag4		What is Player 3's hot dog reciprocity rate? (slider 0–100%)
sag5		What is Player 3's lobster reciprocity rate? (slider 0–100%)

TABLE 7.6
Factors Corresponding to User Metrics

Factor	Description	MR α	DD α
Trust propensity (tp)	The participant's propensity to trust the agent's recommendations.	0.92	0.91
Cognitive reflection (crt)	A measurement of decision-making ability.	0.79	0.73
Reported expertise (re)	The participant's self-assessed domain knowledge.	0.82	0.80
Perceived control (pc)	The participant's subjective assessment of their degree of control over the agent.	0.86	0.96
Perceived transparency (pt)	The participant's subjective assessment of the agent's ability to explain itself.	0.61	0.44
Perceived accuracy (pa)	The participant's subjective assessment of the agent's accuracy.	0.91	0.90
Trust (t)	The participant's reported overall trust in the agent.	0.93	0.90
User experience with the agent (UXA) (ux)	A combination of items from pc, pt, pa, and t.	0.89	0.95

(*Continued*)

TABLE 7.6 (Continued)
Factors Corresponding to User Metrics

Factor	Description	MR α	DD α
Cognitive load (cl)	The participant's subjective assessment of frustration that occurred during the task.	0.82	0.75
Baseline satisfaction (bs)	The participant's self-reported satisfaction with random items (MR only).	0.93	
Decision satisfaction (ds)	The participant's self-reported decision satisfaction (MR only).	0.93	

Note: α is Cronbach's alpha—a measure of internal reliability (this would mean the items that make up the factor are highly correlated). Items that were removed due to poor fit are not shown.

TABLE 7.7
Observed Dependent Variables in the Movie Recommendation Study

Variable Name	Description	MR μ	MR σ	DD μ	DD σ
Recommender int.	Number of interactions with the agent	14	29	25	34
Browser int.	Number of interactions with the simple, alternative tool	37	23		
Adherence	Proportion of the recommendations used in decision making	0.67	0.36	0.33	0.25
Domain knowledge	Score on initial insight questionnaire	0.45	0.16	0.73	0.15
SAT	Score on recommender beliefs questionnaire	0.67	0.2	0.55	0.18
SAG	Score on game state estimation questionnaire			0.99	0.556
Decision optimality	Percentage of moves made that were optimal (0.0 to 1.0)			0.59	0.12

Note: Scores on tests were normalized.

each participant's inherent ease of satisfaction and represents the quantitative change from that level of satisfaction and satisfaction arising from different task factors. The data here indicates that users who would be most likely to take recommendations at face value (without further interaction or investigation) would also be more easily satisfied by a random selection of items. Moreover, it is the knowledgeable users who engage with the system (as evidenced by increased recommender interaction) and understand it (as indicated by increased SAT) that are able to do better, especially when the system allows the user to override its behavior (as evidenced by the effect of Control on Δ Decision Satisfaction).

We believe the results in this work help to demonstrate the value of domain knowledge measurement, SAT, and CRT tests for recommender systems research. These constructs significantly increased the amount of explainable variance in decision satisfaction and adherence without affecting the order of complexity of the regression model. Moreover, their correlations with the user experience construct were quite

low. Given that there were high correlations between the different system perception constructs (control, transparency, accuracy) in this experiment, it might be advisable to reduce the number of subjective user experience questionnaire items and instead use participant time to assess cognitive and knowledge variables.

Many findings in this experiment would have been missed if these measures had been omitted. Users with correct beliefs about the recommender were more likely to adopt recommendations (Figure 7.6). SAT had the highest direct positive impact on adherence with a β coefficient of 0.23, followed closely by the presence of control. User experience did not predict adherence nearly as well as the SAT factor and the control treatment. Furthermore, the "perceived transparency" subconstruct was not nearly as effective at explaining adherence (the tested relationship was nonsignificant in all models). This highlights the need for the use of the objective SAT measure, instead of perceived transparency, within recommender systems research. When combined with its impact on Δ Decision Satisfaction, it highlights the need for recommender system designers to instill deep understanding of recommender operations to maximize engagement, usage, and outcomes.

Increased interaction with the browser side of the interface was linked to increased SAT but also to decreased adherence. To explain this, we examined browser interaction in more detail. We found that, similar to the recommendation side, 50% of browser interactions were filter/sort/search actions and the other 50% were rating actions. What this might suggest is that participants were using the browser tool to find representative movies for their profile. As the participant found more representative items, there was more opportunity to get dynamic feedback from the recommender. Over time, this improved SAT but also increased the chance that the participant found satisfactory items from the browser tool (interesting titles were likely adjacent in metadata space to the targeted titles).

Explanation, control, and recommendation error steered the decision system towards different outcomes. Explanation improved SAT to a degree, which in turn correlated with increased adoption of recommendations and better decision outcomes. However, explanation also nullified the positive effects of control if both were switched on (see Table 7.8); this effect is difficult to explain, because there is no corresponding increase in cognitive load, decrease in user experience, or decrease in

TABLE 7.8
Regressions in the Raykov Change Model That Identifies Factors That Contributed to Improved Decision Making in the Movie Recommendation Task

Regressand	Regression (←)	Coeff.	P(> |z|)
	← UXA	–0.124	*
Δ Decision Satisfaction	← Cognitive Load	–0.237	***
$R^2 = 0.10$	← SAT	0.127	**
	← Control	0.295	***
	← Explanation × Control	–0.268	*
	← Adherence	–0.110	*

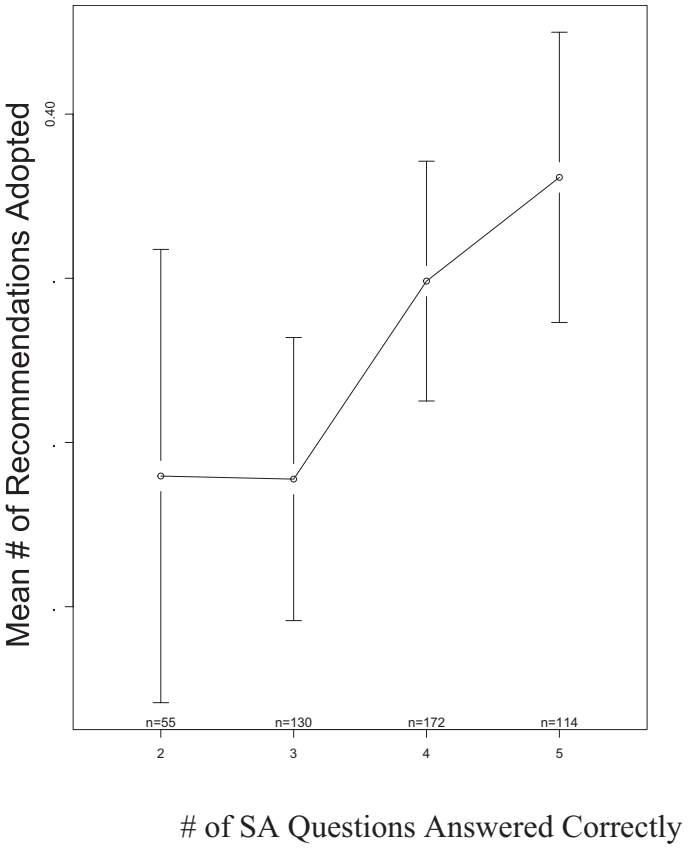

FIGURE 7.6 Relationship between adherence and understanding of the recommender in the Movie Recommendation study.

Note: As understanding increases, users adopt more recommendations.

recommender interaction caused by this particular configuration. A possible reason for this is that explanations boosted confidence in the recommendations, increasing adherence through improved SAT, and thus disengaging the participant. Next, we found that control played two roles. First, control (predictably) increased recommender interaction, which in turn correlated with increased cognitive load and adherence. Second, the presence of control features increased satisfaction with selected items regardless of interaction quantity. This leads us to believe that the ability to have control over a recommender system, whether or not that control is exercised, is a desirable feature of a recommender system because it leads users to be more satisfied with choices. Finally, reductions in recommendation error had the largest impact on user experience but had no direct effect on decision satisfaction. Since an alternative to the recommender was available in this task, it is likely that users switched to the browsing tool when the recommender failed to produce satisfactory results. Our

data also indicates that control has a bigger impact on the user's satisfaction with his/her final watchlist rather than the accuracy of the recommender.

Diner's Dilemma: Statistical Effects

As with the MR study, we found that modeling the individual system perceptions was of little value, but for different reasons. While the perceived control construct was found to be externally discriminant from trust and perceived accuracy, the resulting construct was not predictive in the context of the rest of the model. This is because the alternative model containing a fit perceived control factor not only fit worse than our chosen model, but also complicated the story of reduced interaction caused by the control feature of the Dining Guru.

Perceived transparency had similar problems, but the factor also fit worse and was not predictive. Thus, we chose to model UXA instead, combining the items from trust and perceived accuracy. Also similar to the MR study, participants that considered themselves experts were much less likely to interact with the Dining Guru and adhere to recommendations. Unfortunately, these participants, who reported being trusting and yet performed worse on the domain knowledge test, subsequently ended up scoring fewer points in the game. Simultaneously, these users reported more trust than average with the Dining Guru but less interaction and adherence. Meanwhile, less trusting users demonstrated higher domain knowledge, which predicted more correct beliefs about the recommender and thus more adherence. This situation is strikingly similar to the MR task, but it is not yet clear how these users might be accommodated.

When both explanations and error were present, the model predicts that decision optimality drops below the mean. This is demonstrated in Figure 7.7. This indicates that explanations allowed users to better detect the errors in the Dining Guru, which may have steered them away from adherence in the error-prone treatments. Despite this, adhering to the Dining Guru in even full error condition would have put the user's performance at the mean, and adhering in the weak noise conditions would have put the users well above the mean (recall Table 7.2). This result implies that even relatively accurate decision support systems can be ignored if users are able to detect errors, regardless of the severity, in the agent.

The results from the DD data indicate many positive benefits of incorporating explanations into decision support systems. Explanation indirectly caused increased adherence through SAT and recommender interaction. It also had a direct effect on decision optimality, suggesting that the explanations were useful for helping the participant understand the game. Previously, explanations have been noted to increase trust (Tintarev & Masthoff, 2011), adherence (Arnold et al., 2006), and perceived control (Knijnenburg, Bostandjiev, O'Donovan, & Kobsa, 2012b). Now, this study demonstrates that explanations increase adherence through a mediating effect of the participants' beliefs about the recommender. Additionally, we have observed many important interaction effects between explanation and control or explanation and error. The results suggest that in some situations, explanation and control given together may overload users of the system. Furthermore, explanation features may draw attention to flaws in the system predictions, mitigating automation bias.

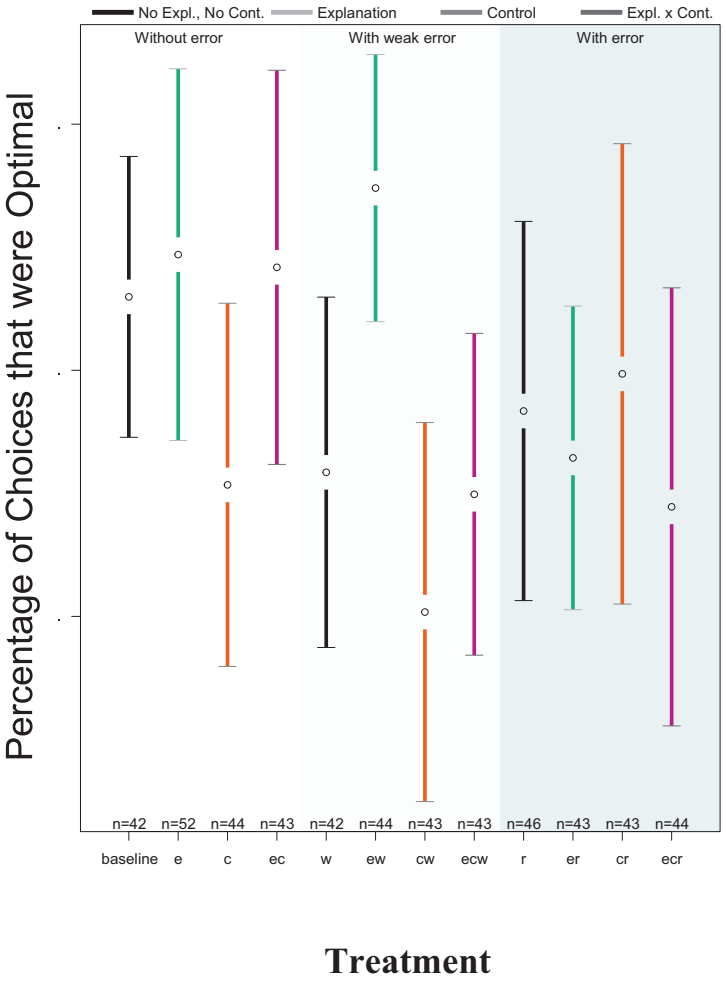

FIGURE 7.7 Mean percent of optimal choices made in each treatment in the Diner's Dilemma study.

Note: Explanations improve performance between "c" and "ec" as well as between "w" and "ew." Error bars are 95% confidence interval.

In some situations, this may lead to better decision making due to the rejection of incorrect predictions.

Comparative Analysis

In this section, the results from both the Movie Recommendation (MR) and Diner's Dilemma (DD) studies are compared. We highlight which results were replicated across both studies. A summary of effects linked to personal user characteristics is shown in Table 7.9. Across both studies, trust propensity predicted

TABLE 7.9
Support for Effects Related to User Profiling Factors

Effect	MR	DD
Trust propensity predicts higher perceptions of an agent	**Yes (***)**	**Yes (**)**
Trust propensity predicts more incorrect beliefs about agent	Yes (**)	No
Cognitive reflection predicts more correct beliefs about agent	Yes (**)	No
Cognitive reflection significantly correlates with domain knowledge	**Yes (***)**	**Yes (***)**
Self-reported expertise predicts less agent interaction	**Yes (***)**	**Yes (***)**
Domain knowledge predicts more agent interaction	Yes (***)	No
Correct beliefs about an agent significantly correlates with trust	No	Yes (**)

Note: Results supported by both studies are shown in bold.

greater perception of the decision support system. CRT also covaried significantly with initial insight tests regardless of domain (in fact, results from both studies suggest humans can be split into high CRT/high knowledge and high trust/high "reported expertise" groups). This suggests the Dunning-Kruger effect (Kruger & Dunning, 1999) is an important cognitive factor to consider when designing human-agent systems. In the MR study, there was a link between trust propensity, user experience, and low SAT, but this was not seen in the DD study, which indicated a covariance between SAT and trust. A link between trust propensity and recommender perceptions was also reported in Knijnenburg et al. (2012a). The agent present in the DD study was relatively simple when compared with the agent from the MR study, which may explain this discrepancy. Finally, users of higher domain knowledge in the MR study interacted more with the recommender, but domain knowledge was not a predictive factor in interaction with the Dining Guru. This may be explained by differences in each agent's facilities: the collaborative filtering algorithm provided information (the recommendation score) that was not present on the browser side of the interface, but the Dining Guru only aided in summarizing information that was already available, perhaps making it less useful to more capable players.

A summary of effects for the participant's cognitive states are shown in Table 7.10. Across both studies, SAT was an effective mediator of the effects of explanation on adherence and was also predictive of decision outcomes. This extends our understanding of the importance of explanations past the subjective realm of system perceptions and trust. Moreover, the positive effects on decision outcomes provide quantitative data to suggest that accurate but incomprehensible systems (e.g., deep learning) may require additional research in transparency. Where decision success was objective (DD), higher domain knowledge directly predicted better decision performance. Finally, we found that when controlling for domain knowledge in the DD study, CRT was an unnecessary predictor. This is likely due to limitations in our knowledge test for the MR study, which may have been less effective at capturing task concepts, or simply because Δ Decision Satisfaction in the MR study was independent of each participant's knowledge.

TABLE 7.10
Support for Effects Related to SAT and Domain Knowledge

Effect	MR	DD
Correct beliefs about an agent predicts increased adherence	**Yes (***)**	**Yes (*)**
Domain knowledge predicts better decision performance		Yes (***)
Correct beliefs about an agent predicts better decision performance	**Yes (**)**	**Yes (***)**
Cognitive load and user experience with an agent are negatively correlated	**Yes (***)**	**Yes (***)**
Higher user experience with an agent predicts worse decision performance	Yes (*)	No
Higher user experience with an agent predicts adherence	Yes (**)	No
More interaction with the agent predicts increased adherence	**Yes (***)**	**Yes (***)**

Note: Results supported by both studies are shown in bold.

TABLE 7.11
Support for Effects Caused by Altering the Agent's ECR Profile

Effect	MR	DD
Explanation causes correct beliefs about an agent	**Yes (*)**	**Yes (***)**
Explanation causes improved decision outcomes	**Yes (*, full mediation via SAT)**	**Yes (**)**
Control causes increased cognitive load	**Yes (**)**	**With explanation (**)**
Control causes increased adherence	Yes (***, full mediation via interaction)	No
Control increases decision performance	Yes (***)	No
Error decreases user experience with an agent	Yes (***)	No
Error decreases decision performance	No	With explanation (**)
Error decreases adherence	**Yes (**, full mediation)**	**Yes (**)**

Note: Results supported by both studies are shown in bold.

Across both studies, cognitive load was negatively correlated with user perceptions of the agent, indicating the potential for an agent to mentally relieve analysts. Higher user experience with the agent only led to increased adherence in the MR study. Moreover, higher UXA was linked to higher satisfaction with selected items in the MS study. As previously discussed, this satisfaction decrease was linked to lower engagement with the system, suggesting a need for maximizing interaction for the best outcome.

A summary of claims on explanation, control, and error of an agent is shown in Table 7.11. In both studies, the presence of explanations caused better decision success and better agent understanding. Recall that in both studies, explanation was given under varying levels of agent error. In the MR study, adherence dropped slightly when the agent made errors, but overall decision outcomes were not affected. In the DD study, error simultaneously predicts increased interaction with the agent, but also decreased adherence. When explanations were given alongside the erroneous

recommendations, overall decision outcomes suffered. These results suggest that agent errors lead to complex situations. Explanations can potentially help users identify when an agent makes errors so that alternatives can be used instead. However, users may under-trust the system (DD) when, despite making errors, the average performance is still higher than the human operators. Moreover, if adequate alternative systems are accessible (MR), errors may not make their way into the final decision outcome, but will likely reflect on adherence to the agent, resulting in disuse.

Control features increased cognitive load across both the MR and DD studies. Control features in the MR study allowed users to customize the recommendation view to their tastes, getting the benefits of both traditional filtering and collaborative filtering. Control features in the DD study allowed users to explore the space of decision outcomes and also the "automation" from the Dining Guru. Users adhered less to the Dining Guru's recommendations when given control, due to decreased interaction, and thus decision optimality suffered (again, the Dining Guru performed significantly better than the mean in most treatments). Likewise, the control feature in the MR study gave the participant an increased ability to explore the movie catalog space, resulting in better Δ Decision Satisfaction. The negative outcomes associated with control in the DD study may have been due to a usability issue: the system required significantly more effort to use over the automated version, and unlike the explanation feature, there was no evidence that the control feature helped participants understand the game or the agent better. While the control features were designed to be analogous for their respective tasks, this aspect of the agent appears to be the most sensitive. Moreover, cognitive load appears to be a reliably unfortunate side effect of adding control features. We suggest that agent designers carefully consider and iteratively prototype control features to complement the task domain and agent.

Finally, agent errors predicted decreased adherence in both studies. In the MR study, this effect was fully mediated by user perception (with no direct effect found), indicating that users may have simply turned to the browser side of the interface when the recommender failed. In the DD study, the negative effects of error were partially mediated by recommender interaction, indicating only a minor decrease in adherence. As mentioned previously, this had a negative effect on participant performance, which was unfortunate. This analysis suggests that users may be overly sensitive to *perceived* errors on the part of an agent, which may be exacerbated when the user is overconfident (high reported expertise). In multi-trial tasks, perhaps agents can convince their users with a retrospective "if you had followed my advice . . ." argument. This is a promising area for future work.

Extending to Multi-Agent Systems

This chapter has presented two studies that have taken a step in quantitatively mapping out cognitive and behavioral factors for monolithic non-embodied agents (technological artifacts). The area of multi-agent systems is much less explored, particularly in the area of interacting with multi-agent swarms. This section concludes with a discussion of how the cognitive factors presented here would be relevant in multi-agent systems and swarms.

Multi-agent systems generally operate in objective task domains, where there is a tangible task such as surveillance that must be completed. The distribution of the

agents is made to augment human decision-making capabilities according to concrete guidelines. Thus, the findings of the DD study will most strongly inform the discussion of this section. On occasion, multi-agent systems can be used to model the subjective preferences of humans to study their interactions. This is especially true in studies for marketing and crowd control (Humann & Madni, 2014; Kadyrova & Panasyuk, 2016). In these cases, the agents act as surrogates for human decision makers, in order to simulate the effects of system designs on the public.

With multiple interacting tangible agents, there is a high risk of cognitive overload. The greatest difficulty arises when interactions among agents are essential for the performance of the systems. In this case, the potential interactions grow quadratically with the number of agents in the system, quickly overwhelming a human's ability to track them all. Even just the sight of groups of interacting robots is enough to cause elevated stress levels in test subjects (Podevijn et al., 2016). Therefore, detailed explanation of an individual agent's decision process is almost always avoided. Instead, it may be necessary to generate monolithic explanations for multi-agent systems to reap the SAT and trust-related benefits. Moreover, multi-agent systems are often automated when gathering and summarizing data before it is presented to the user in raw form, reducing explanation. Repeated requests for attention from agents can quickly annoy and overwhelm a human controller.

Much research has gone into finding the fan-out of a human (i.e. the number of robots that one human can control) (Crandall, Goodrich, Olsen, & Nielsen, 2005; Humann & Pollard, 2019). A human must from time to time switch his attention among robots, and this imposes a cost in both time and situational awareness (Goodrich, Quigley, & Cosenzo, 2005). As a rough estimate, the fan-out of an operator can be calculated from the task switching metrics *interaction time* and *neglect time* (Chen, Barnes, & Harper-Sciarini, 2011b; Crandall et al., 2005). In the emerging field of swarm robotics, where the number of agents can reach into the thousands, there is no hope for human comprehension or control, or even mathematical prediction of behavior (Edmonds, 2004), so self-organizing algorithms, simulation (Humann, Khani, & Jin, 2014), and statistical testing are used to gain confidence in system performance.

The single-agent studies in this chapter show no strong correlation between cognitive load and decision optimality, but this may be because the users were not pushed to their cognitive limits by the monolithic agent. A multi-agent system which is prompting the user for control and decision making under time constraints represents a much more cognitively taxing scenario and may show a stronger relationship if it were studied in the same way.

Situation awareness is a design challenge both in the multi-agent system itself and in the user interface. Because most multi-agent systems are focused on objective tasks, global SA (analogous to SAG in the DD study) is the primary concern. Global SA is especially important when a human is used as a backup or troubleshooter for an automated system. In these cases, the automation is used for efficiency and precision during the majority of task execution, but when errors arise, the human is called in to restore functionality. If the human is not actively engaged with the system and maintaining global SA, it can be difficult to make accurate quick decisions when abruptly called back into duty (Ordoukhanian & Madni, 2017, 2018). Attempting to

maintain SAT of large multi-agent systems is often irrelevant and could be harmful if it demands too much of the user's attention. The results of the DD study do not show an interaction between SAT and SAG, but we can speculate that as more cognitive effort is expended tracking the states of the individual agents, the user will have less ability to maintain awareness of the task and total system state. Therefore, global SA can be aided by design of low-explanation UIs that only present relevant task-centric information at the expense of more detailed SAT. Comprehension of the significance of system states could also be aided by UIs that alert the user when certain critical states are reached.

Projecting how multi-agent systems will behave in the future is perhaps the most critical research aspect, as this is an area where humans tend to falter (Tabibian et al., 2014). Systems can be made more predictable when designed with the human in mind, or UIs can present predictions directly to the user. If predictions are presented, it would be necessary to provide a limited volume of explanation, as they are based on simulations whose underpinnings would be difficult for the user to comprehend in real time.

Over-reliance on automation is even more of a risk in multi-agent systems, as the systems can quickly become so incomprehensible that the user's only viable choice to remain in control is to naively trust the recommendations of the expert system (Parasuraman & Riley, 1997). While trust propensity is only indirectly linked to adherence in these studies, it may need to be a part of the user's training with the system. Recommenders within multi-agent systems are built with the assumption that they can provide more optimal functionality than human judgment alone, so users may need to be taught to trust the system even in the face of errors and stress.

Unlike in monolithic systems, people may form a perception of each individual agent, so that the system perception may not be simply binary, as is indicated by Hassenzahl et al. (2008; Hassenzahl & Tractinsky, 2006). Here the concept of trust becomes more complicated, as a user may have differing levels of trust in each agent, and these may not all be easily predicted from that operator's trust propensity. Although control was not shown to strongly effect UXA, we can speculate that in a multi-agent system, having the ability to change or quarantine untrustworthy agents would have a positive effect on UXA. (But again, this may be counterproductive if the user substitutes his own erroneous judgment because he mistakenly distrusts an agent.)

Design of Intelligent Multi-Agent Systems and Future Work

The benefits and of multi-agent systems have made them an attractive design goal, but their complexity makes design and use very challenging. Traditional design processes can falter because of the unpredictability of multi-agent behavior, so advanced modeling, simulation, and optimization are often needed (Humann, 2015). These techniques include multi-objective and hierarchical optimization (Durand, Burgaud, Cooksey, & Mavris, 2017; Fisher, Cooksey, & Mavris, 2017), genetic algorithms to tune agent behavioral parameters (Humann & Jin, 2013; Humann et al., 2014), multi-agent simulation (including simulation of human behavior) (Gao & Cummings, 2012), and many others. The common theme is that computational tools are necessary to aid designers, as the complexity of systems is too great to draw them up fully formed from intuition or analytical methods.

Future multi-agent systems will require a more thorough understanding of human factors. This includes discovering the limitations of humans from the perspective of the agents. Agents will need to be able to sense when a human is overloaded or performing poorly and adapt. Executable models of humans will need to be developed in control and decision-making scenarios, so that designers can use realistic simulations of humans to test system autonomy levels prior to manufacture and deployment. Universal metrics for human interaction with multi-agent systems must be developed so that results of case studies can be interpreted and applied to new ideas.

Finally, emerging technology for human-systems interaction will also need to be deployed. This includes augmented reality, which can be used to summarize highlights of swarm states for situation awareness, and multi-modal (e.g. haptic) feedback to keep the user informed of the system status across multiple channels without overloading a single channel.

To reach these goals, we propose the following research agenda:

- Study the relationship between global SA and SAT: most UIs for multi-agent systems are designed under the assumption that users should be shielded from the internal details of each agent, only focusing on high-level tasks and system states. Is this assumption sound? Can a more detailed knowledge of SAT lead to inference of global SA?
- Pinpoint limits of human cognitive load: multi-agent environments will be much more cognitively taxing on the user, especially if he is forced to make decisions under time constraints. Effective design of systems must take the user's limits into account. How can workload be predicted by designers? How is workload related to decision optimality?
- Clarify relationships between levels of situation awareness and cognitive load: the "levels" of situation awareness (knowledge of states, comprehension of states, and prediction of future states) may not strictly build on one another, and any one could be tracked and summarized by a UI. Does attempting to maintain these different levels have a different effect on cognitive load? If one or more are presented to the user through a UI, making them easier to maintain, how does this affect cognitive load?
- Investigate forced vs. voluntary users: many multi-agent systems are designed out of necessity; it is unrealistic for users to complete the task without the aid of the system. In a work environment, employees can be forced to use the system. How does this affect trust in the system? Can users be trained to be more trusting? How does it affect UXA, and in the end does UXA only matter when trying to attract voluntary users?

Summary

This chapter has investigated how human cognition reacts to the presence and configuration of monolithic agents. We have identified general system, user, and cognitive factors that predict decision behaviors related to interaction with systems, incorporate of system predictions (adherence), and domain decision success. We have presented surprising effects related to user traits and beliefs about systems

that opens a door to future investigations. The analysis of multiple domains and the use of a common measurement methodology in two experiments has allowed us to better identify effects that should generalize well to other contexts. Furthermore, we have discussed these effects in the context of multi-agent systems and identified future research and challenges.

In the introduction, we posed the following research questions:

- How do a person's cognitive traits affect usage of an intelligent agent and resulting decision outcomes?
- Which cognitive or system factors explain variability in decision making (interaction, adherence, success) in the HAI system?
- What is the relationship between correct beliefs about agents, their use, and trust?

We provide the following answers to these research questions.

(1) In this work, we have quantitative evidence that suggests that self-reported experts are likely to be more trusting than the general population of users. These users not only interact less but also adhere to advice more often. True domain experts are more likely to have higher CRT. Intelligent agents could potentially adapt to users based on their personality, however, how to accommodate overconfident users remains an open research question.
(2) Despite the subjective (MR) vs. objective (DD) parameters of each task, we found that user experience and cognitive load were not as important as a user's understanding of task factors or understanding of the agent. Manipulation of the system's ECR profile also more strongly affected outcomes than subjective system perceptions.
(3) Situation-awareness based agent transparency (SAT) was an effective mediator of system explanation effects when trying to understand adherence to advice. Furthermore, there is strong evidence here to suggest that SAT and trust/system perceptions are discriminant, while SAT was found to be externally valid. This suggests that less trusting users might be convinced to use a system through effecting correct beliefs.

While this research has identified a number of HAI factors that transfer across domains and while we have provided expectations for their general relationships in a very limited scope, more research in other decision and task contexts, especially multi-agent tasks, is needed to develop a reliable, general theory about how intelligent agents affect human cognition and decision-making behavior. Additional factor modeling, especially task- and domain-specific factors, will be essential in achieving high levels of prediction about how human-machine systems evolve. This study has also not examined the longitudinal effects of repeated agent use on cognitive factors, nor how relationships between users and agents evolve over long periods of time. The domain knowledge and SAT metrics used in this task are exploratory and require further validation and study in each task domain where they are applied. Finally, the effects reported here warrant further and more detailed study where more variables are controlled.

In summary, we have discovered that cognitive traits and intermediate cognitive variables are crucial for understanding the effects of explanation, control, and error for HAI. Furthermore, we discovered that (1) the user profiling metrics trust propensity, cognitive reflection, reported knowledge, and domain knowledge increase the ability to predict decision-making behaviors in the presence of an agent, (2) objectively defined metrics such as situation awareness and domain knowledge are more indicative of outcomes than subjective system perceptions in the HAI system, and (3) correct user beliefs (SAT) about an agent mediate the effect of system explanation when predicting adherence to recommendations.

NOTES

All URLs last accessed June 2020.

1. www.vanityfair.com/news/business/2014/10/air-france-flight-447-crash
2. Our approach was nearly identical to http://grouplens.org/blog/similarity-functions-for-user-user-collaborative-filtering/
3. www.qualtrics.com/
4. www.wired.com/2012/10/lance-armstrong-and-the-prisoners-dilemma-of-doping-in-professional-sports/
5. http://www.usabart.nl/QRMS/

REFERENCES

Ahuja, R. K., Mehlhorn, K., Orlin, J., & Tarjan, R. E. (1990). Faster algorithms for the shortest path problem. *Journal of the ACM (JACM)*, *37*(2), 213–223.

Ainslie, G. (2001). *Breakdown of will*. Cambridge: Cambridge University Press.

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

Bainbridge, W. A., Hart, J., Kim, E. S., & Scassellati, B. (2008). The effect of presence on human-robot interaction. In *Robot and human interactive communication, 2008. RO-MAN 2008. The 17th IEEE international symposium on* (pp. 701–706). New York: IEEE.

Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems*, *6*(3), 4.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289–300.

Bilgic, M., & Mooney, R. J. (2005). Explaining recommendations: Satisfaction vs. promotion. In *Beyond personalization workshop*, *IUI* (Vol. 5).

Breazeal, C. (2004). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *34*(2), 181–186.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the fourteenth conference on uncertainty in artificial intelligence* (pp. 43–52). Burlington, MA: Morgan Kaufmann Publishers Inc.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Cafaro, A., Vilhjálmsson, H. H., Bickmore, T., Heylen, D., Jóhannsdóttir, K. R., & Valgarðsson, G. S. (2012). First impressions: Users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *International conference on intelligent virtual agents* (pp. 67–80). London: Springer.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Chen, J. Y., & Barnes, M. J. (2014). Human—agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, *44*(1), 13–29.

Chen, J. Y., Barnes, M. J., & Harper-Sciarini, M. (2011a). Supervisory control of multiple robots: Human-performance issues and user-interface design. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, *41*(4), 435–454.

Chen, J. Y., Barnes, M. J., & Harper-Sciarini, M. (2011b). Supervisory control of multiple robots: Human-performance issues and user-interface design. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, *41*(4), 435–454.

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). Situation awareness-based agent transparency. *Technical report*, DTIC Document.

Chi, E. H. (2015). Blurring of the boundary between interactive search and recommendation. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 2–2). New York: ACM.

Choi, A., de Melo, C. M., Khooshabeh, P., Woo, W., & Gratch, J. (2015). Physiological evidence for a dual process model of the social effects of emotion in computers. *International Journal of Human-Computer Studies*, *74*, 41–53.

Choi, S., & Clark, R. E. (2006). Cognitive and affective benefits of an animated pedagogical agent for learning English as a second language. *Journal of Educational Computing Research*, *34*(4), 441–466.

Colquitt, J. A., Scott, B. A., & LePine, J. A. (2007). Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance. *Journal of Applied Psychology*, *92*(4), 909.

Crandall, J. W., Goodrich, M. A., Olsen, D. R., & Nielsen, C. W. (2005). Validating human-robot interaction schemes in multitasking environments. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *35*(4), 438–449.

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

Cummings, M. (2004). Automation bias in intelligent time critical decision support systems. In *AIAA 1st intelligent systems technical conference* (p. 6313). Reston, VA: AIAA.

Davidson, J. E., & Downing, C. (2000). Contemporary models of intelligence. In *Handbook of intelligence* (pp. 34–49). Cambridge, UK: Cambridge University Press.

Durand, J.-G. J., Burgaud, F., Cooksey, K. D., & Mavris, D. N. (2017). A design optimization technique for multi-robot systems. In *55th AIAA Aerospace sciences meeting* (p. 0690). Reston, VA: AIAA.

Edmonds, B. (2004). Using the experimental method to produce reliable self-organised systems. In *International workshop on engineering self-organising applications* (pp. 84–99). London: Springer.

Endsley, M. R. (1995a). Measurement of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *37*(1), 65–84.

Endsley, M. R. (1995b). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32–64.

Endsley, M. R. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. In *Situation awareness analysis and measurement* (p. 10). Boca Raton, FL: CRC Press.

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The Information Society*, *20*(5), 325–344.

Fisher, Z. C., Cooksey, K. D., & Mavris, D. (2017). A model-based systems engineering approach to design automation of SUAS. In *Aerospace conference, 2017 IEEE* (pp. 1–15). New York: IEEE.

Gao, F., & Cummings, M. (2012). Using discrete event simulation to model multi-robot multi-operator teamwork. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 56, pp. 2093–2097). Los Angeles, CA: Sage Publications.

Gill, H., Boies, K., Finegan, J. E., & McNally, J. (2005). Antecedents of trust: Establishing a boundary condition for the relation between propensity to trust and intention to trust. *Journal of Business and Psychology*, *19*(3), 287–302.

Gomez-Uribe, C. A., & Hunt, N. (2016). The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, *6*(4), 13.

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

Goodrich, M. A., Quigley, M., & Cosenzo, K. (2005). Task switching and multi-robot teams. In *Multi-robot systems: From swarms to intelligent automata* (Vol. III, pp. 185–195). Netherlands: Springer.

Gregersen, H., & Sailer, L. (1993). Chaos theory and its implications for social science research. *Human Relations*, *46*(7), 777–802.

Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 497–530.

Gunning, D. (2017). Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*, nd Web.

Harman, J. L., O'Donovan, J., Abdelzaher, T., & Gonzalez, C. (2014). Dynamics of human trust in recommender systems. In *Proceedings of the 8th ACM conference on recommender systems* (pp. 305–308). New York: ACM.

Harper, F. M., & Konstan, J. A. (2016). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *5*(4), 19.

Hassenzahl, M. (2008). User experience (UX): Towards an experiential perspective on product quality. In *Proceedings of the 20th conference on l'Interaction HommeMachine* (pp. 11–15). New York: ACM.

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & Information Technology*, *25*(2), 91–97.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. In *Behavior Research Methods* (Vol. 48, No. 1, pp. 400–407). Cham, Switzerland: Springer Nature.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of ACM CSCW'00 conference on computer-supported cooperative work* (pp. 241–250). New York: ACM.

Hertzum, M., Andersen, H. H., Andersen, V., & Hansen, C. B. (2002). Trust in information sources: seeking information from people, documents, and virtual agents. *Interacting with Computers*, *14*(5), 575–599.

Holmes, G., Donkin, A., & Witten, I. H. (1994). Weka: A machine learning workbench. In *Intelligent information systems, 1994: Proceedings of the 1994 second Australian and New Zealand conference on* (pp. 357–361). New York: IEEE.

Hoorens, V. (1993). Self-enhancement and superiority biases in social comparison. *European Review of Social Psychology*, *4*(1), 113–139.

Humann, J. (2015). *Behavioral modeling and computational synthesis of self-organizing systems*. Doctoral Thesis, University of Southern California, ProQuest, Ann Arbor, MI.

Humann, J., & Jin, Y. (2013). Evolutionary design of cellular self-organizing systems. In *ASME 2013 international design engineering technical conferences and computers and information in engineering conference* (pp. V03AT03A046–V03AT03A046). New York: American Society of Mechanical Engineers.

Humann, J., Khani, N., & Jin, Y. (2014). Evolutionary computational synthesis of self-organizing systems. *AI EDAM*, *28*(3), 259–275.

Humann, J., Khani, N., & Jin, Y. (2016). Adaptability tradeoffs in the design of self-organizing systems. In *ASME 2016 international design engineering technical conferences and computers and information in engineering conference* (pp. V007T06A016–V007T06A016). New York: American Society of Mechanical Engineers.

Humann, J., & Madni, A. M. (2014). Integrated agent-based modeling and optimization in complex systems analysis. *Procedia Computer Science*, *28*, 818–827.

Humann, J., & Pollard, K. A. (2019). Human factors in the scalability of multirobot operation: A review and simulation. In *2019 IEEE international conference on Systems, Man and Cybernetics (SMC)* (pp. 700–707). New York: IEEE.

Humann, J., & Spero, E. (2018). Modeling and simulation of multi-UAV, multi-operator surveillance systems. In *Systems conference (SysCon), 2018 annual IEEE international* (pp. 1–8). New York: IEEE.

Jin, Y., & Levitt, R. E. (1996). The virtual design team: A computational model of project organizations. *Computational Mathematical Organization Theory*, *2*(3), 171–195.

Jung, J. J. (2012). Attribute selection-based recommendation framework for shorthead user group: An empirical study by MovieLens and IMDB. *Expert Systems with Applications*, *39*(4), 4049–4054.

Kadyrova, L., & Panasyuk, M. (2016). Simulation modeling of consumer behavior in decision making about point of services purchase. *Academy of Marketing Studies Journal*, *20*, 70.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., & Kobsa, A. (2012a). Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on recommender systems* (pp. 43–50). New York: ACM.

Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., & Kobsa, A. (2012b). Inspectability and control in social recommenders. In P. Cunningham, N. J. Hurley, I. Guy, & S. S. Anand (Eds.), *RecSys* (pp. 43–50). New York: ACM.

Knijnenburg, B. P., Willemsen, M. C., Gantner, Z., Soncu, H., & Newell, C. (2012). Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4–5), 441–504.

Komiak, S. Y., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quarterly*, 941–960.

Koren, Y., & Bell, R. (2015). Advances in collaborative filtering. In *Recommender systems handbook* (pp. 77–118). London: Springer.

Kruger, J. (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), 221.

Kruger, J., & Dunning, D. (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), 1121.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2017). Interpretable & explorable approximations of black box models. In *Proceedings of the KDD workshop on fairness, accountability, and transparency in machine learning*. New York: ACM.

Lee, M. K., & Turban, E. (2001). A trust model for consumer internet shopping. *International Journal of Electronic Commerce*, *6*(1), 75–91.

Marik, V., & McFarlane, D. (2005). Industrial adoption of agent-based technologies. *IEEE Intelligent Systems*, *20*(1), 27–35.

Martin, J. M., Juvina, I., Lebiere, C., & Gonzalez, C. (2011). The effects of individual and context on aggression in repeated social interaction. In *Engineering psychology and cognitive ergonomics* (pp. 442–451). London: Springer.

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)*, *2*(2), 12.

McKnight, H., Carter, M., & Clay, P. (2009). Trust in technology: Development of a set of constructs and measures. In *Digit 2009 proceedings* (p. 10). Atlanta, GA: Association for Information Systems.

Merritt, K., Smith, D., & Renzo, J. (2005). An investigation of self-reported computer literacy: Is it reliable. *Issues in Information Systems*, *6*(1), 289–295.

Mikhailov, A. S. (2011). *From swarms to societies: Origins of social organization* (pp. 367–380). London: Springer.

Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., & Riedl, J. (2003). MovieLens unplugged: Experiences with an occasionally connected recommender system. In *Proceedings of the 8th international conference on intelligent user interfaces* (pp. 263–266). New York: ACM.

Norman, D. A. (1986). Cognitive engineering. In *User centered system design: New perspectives on human-computer interaction* (p. 3161). Hillsdale, NJ: Erlbaum.

O'Donovan, J., & Smyth, B. (2005). Trust in recommender systems. In *Proceedings of the 10th international conference on intelligent user interfaces* (pp. 167–174). New York: ACM.

O'Donovan, J., Tintarev, N., Felfernig, A., Brusilovsky, P., Semeraro, G., & Lops, P. (2015). Joint workshop on interfaces and human decision making for recommender systems. In H. Werthner, M. Zanker, J. Golbeck, & G. Semeraro (Eds.), *RecSys* (pp. 347–348). New York: ACM.

Ordoukhanian, E., & Madni, A. M. (2017). Human-systems integration challenges in resilient multi-uav operation. In *International conference on applied human factors and ergonomics* (pp. 131–138). London: Springer.

Ordoukhanian, E., & Madni, A. M. (2018). *Introducing resilience into multi-UAV system-of-systems network* (pp. 27–40). London: Springer.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71.

Panait, L., & Luke, S. (2005). Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-agent Systems*, *11*(3), 387–434.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140–160.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker.* Cambridge: Cambridge University Press.

Podevijn, G., O'grady, R., Mathews, N., Gilles, A., Fantini-Hauwel, C., & Dorigo, M. (2016). Investigating the effect of increasing robot group sizes on the human psychophysiological state in the context of human-swarm interaction. *Swarm Intelligence*, *10*(3), 193–210.

Prokopenko, M. (2013). *Advances in applied self-organizing systems*. London: Springer.

Pu, P., Chen, L., & Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on recommender systems* (pp. 157–164). New York: ACM.

Raykov, T. (1992). Structural models for studying correlates and predictors of change. *Australian Journal of Psychology*, *44*(2), 101–112.

Ren, Z., Yang, F., Bouchlaghem, N., & Anumba, C. (2011). Multi-disciplinary collaborative building design—a comparative study between multi-agent systems and multi-disciplinary optimisation approaches. *Automation in Construction*, *20*(5), 537–549.

Requicha, A. (2013). *Swarms of self-organized nanorobots* (pp. 41–49). London: Springer.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of ACM CSCW'94 conference on computer-supported cooperative work* (pp. 175–186). New York: ACM.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). New York: ACM.

Rosenfeld, A., Agmon, N., Maksimov, O., & Kraus, S. (2017). Intelligent agent supporting human—multi-robot team collaboration. *Artificial Intelligence*, *252*, 211–231.

Rubenstein, M., Ahler, C., & Nagpal, R. (2012). Kilobot: A low cost scalable robot system for collective behaviors. In *Robotics and Automation (ICRA), 2012 IEEE international conference on* (pp. 3293–3298). New York: IEEE.

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (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* (pp. 141–148). New York: ACM.

Schaffer, J., O'Donovan, J., & Höllerer, T. (2018). Easy to please: Separating user experience from choice satisfaction. In *Proceedings of the 26th conference on user modeling, adaptation and personalization* (pp. 177–185). New York: ACM.

Shortliffe, E. H., Davis, R., Axline, S. G., Buchanan, B. G., Green, C. C., & Cohen, S. N. (1975). Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research*, *8*(4), 303–320. Amsterdam: Elsevier.

Sinha, R., & Swearingen, K. (2002). The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems* (pp. 830–831). New York: ACM.

Snyder, M. G., Qu, Z., Chen, J. Y., & Barnes, M. J. (2010). Roboleader for reconnaissance by a team of robotic vehicles. In *Collaborative Technologies and Systems (CTS), 2010 international symposium on* (pp. 522–530). New York: IEEE.

Stephens, D. W., McLinn, C. M., & Stevens, J. R. (2002). Discounting and reciprocity in an iterated prisoner's dilemma. *Science*, *298*(5601), 2216–2218.

Stuart, D., Christensen, K., Chen, A., Cao, K.-C., Zeng, C., & Chen, Y. (2013). A framework for modeling and managing mass pedestrian evacuations involving individuals with disabilities: Networked Segways as mobile sensors and actuators. In *ASME 2013 international design engineering technical conferences and computers and information in engineering conference* (pp. V004T08A011–V004T08A011). New York: American Society of Mechanical Engineers.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312.

Tabibian, B., Lewis, M., Lebiere, C., Chakraborty, N., Sycara, K., Bennati, S., & Oishi, M. (2014). Towards a cognitively-based analytic model of human control of swarms. In *2014 AAAI spring symposium series*. Palo Alto, CA: AAAI.

Tintarev, N., Kang, B., Höllerer, T., & O'Donovan, J. (2015). Inspection mechanisms for community-based content discovery in microblogs. In *IntRS@ RecSys* (pp. 21–28). New York: ACM.

Tintarev, N., & Masthoff, J. (2007). A survey of explanations in recommender systems. In *Data engineering workshop, 2007 IEEE 23rd international conference on* (pp. 801–810). New York: IEEE.

Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for recommender systems. In *Recommender Systems Handbook*, pages 479–510. Springer.

Tintarev, N., O'Donovan, J., Brusilovsky, P., Felfernig, A., Semeraro, G., & Lops, P. (2014). Recsys' 14 joint workshop on interfaces and human decision making for recommender systems. In *Proceedings of the 8th ACM conference on recommender systems* (pp. 383–384). New York: ACM.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289.

Ullman, J. B., & Bentler, P. M. (2003). *Structural equation modeling*. Hoboken, NJ: Wiley Online Library.

Varian, H. R., Bergstrom, T. C., & West, J. E. (1996). *Intermediate microeconomics* (Vol. 4). New York: Norton.

Veletsianos, G. (2007). Cognitive and affective benefits of an animated pedagogical agent: Considering contextual relevance and aesthetics. *Journal of Educational Computing Research*, *36*(4), 373–377.

Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Mataric, M. J. (2006). The role of physical embodiment in human-robot interaction. In *Robot and human interactive communication, 2006. ROMAN 2006. The 15th IEEE international symposium on* (pp. 117–122). New York: IEEE.

Welsh, M., Burns, N., & Delfabbro, P. (2013). The cognitive reflection test: How much more than numerical ability. In *Proceedings of the 35th annual conference of the cognitive science society* (pp. 1587–1592). Cognitive Science Society.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd international conference on intelligent user interfaces* (pp. 307–317).New York: ACM.