The Effect of Displaying System Confidence Information on the Usage of Autonomous Systems for Non-specialist Applications: A Lab Study

Jhim Kiel M. Verame, Enrico Costanza and Sarvapali D. Ramchurn
University of Southampton
Southampton, United Kingdom
{j.verame, e.costanza, sdr1}@soton.ac.uk

ABSTRACT
Autonomous systems are designed to take actions on behalf of users, acting autonomously upon data from sensors or online sources. As such, the design of interaction mechanisms that enable users to understand the operation of autonomous systems and flexibly delegate or regain control is an open challenge for HCI. Against this background, in this paper we report on a lab study designed to investigate whether displaying the confidence of an autonomous system about the quality of its work, which we call its confidence information, can improve user acceptance and interaction with autonomous systems. The results demonstrate that confidence information encourages the usage of the autonomous system we tested, compared to a situation where such information is not available. Furthermore, an additional contribution of our work is the method we employ to study users' incentives to do work in collaboration with the autonomous system. In experiments comparing different incentive strategies, our results indicate that our translation of behavioural economics research methods to HCI can support the study of interactions with autonomous systems in the lab.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI).

Author Keywords
autonomous agent; behavioural economics; lab study; confidence information

INTRODUCTION
Autonomous systems are designed to take actions on behalf of the user, acting autonomously upon data from sensors or online sources. Because of the increasing availability of low-cost sensors, actuators, computational devices and large amounts of online data, in recent years these types of systems are becoming more prevalent around "non-specialist applications", applications where users are not expected to be trained to use them. Practical real-world examples include smart appliances, such as smart thermostats[1], or autonomous software, so called *agents* that can bid for users in on-line auction websites[2].

Generally, autonomous systems are based on techniques such as machine learning and artificial intelligence to process input data (be it from sensors or online sources) and automatically take decisions to guide their autonomous operation. However, because of noise and biases in real world data, limited size of training data sets, discrepancies between computationally feasible models and complex real-life systems, the results of automatic data analysis and classification may often be liable to considerable uncertainty. Therefore, for many practical applications it is important to allow users to easily delegate or regain control based on their expectations about the capabilities of the autonomous system, an idea known as "flexible autonomy" [20]. As a consequence, the design of interaction mechanisms that enable users to understand the operation of autonomous systems and flexibly delegate or regain control is currently an open challenge for HCI [8, 38].

While studies of interaction with autonomous systems for specialist applications (e.g. disaster response or aviation) date back to the 1970s [33], it is only more recently that research has focused on the adoption of autonomous products in the home, such as the Nest thermostat [37, 38]. Findings from these studies suggest that because people find it difficult to recognise how well such products work, they tend to not use them. They become frustrated, so their interaction with such systems decrease over time, which may potentially lead to the abandonment of this technology. Recent work has suggested that the display of *confidence information* can increase user's awareness of the ability of autonomous systems [5, 18]. Confidence information is the estimated probability that an inference produced by a smart system is correct, under the assumption that the system has the correct model to interpret the data[3]. In this paper we report a lab study (N=60) designed to investigate whether displaying confidence information can improve user acceptance and interaction with autonomous systems. The

CHI'16, May 07 – May 12, 2016, San Jose, CA, USA

http://dx.doi.org/10.1145/2858036.2858369

[1] http://www.nest.com/
[2] http://www.snipeswipe.com/
[3] In other words, our work is based on the assumption that the confidence information is reliable. While such assumption is realistic for a number of smart systems, it is worth noting that in some cases incorrect models can produce confidence information that is unreliable.

results demonstrate that confidence information encourages the usage of the autonomous system we tested, compared to situations where such information is not available. Indeed, this is the primary contribution of our work.

Moreover, while recent work on interaction with autonomous systems for non-specialist activities has been based on field studies [1, 8, 37, 38], we are interested in exploring the opportunity to study interaction with such systems through controlled lab studies. Notwithstanding the importance of field trials, we see lab studies as an important complementary research tool. Maintaining high ecological validity is particularly challenging in the design of lab studies; to address it, we look at research methods from the behavioural economics literature. In particular, we demonstrate that by using financial incentives and repeated tasks in the experimental design it is possible to create a situation where participants' decision to use an autonomous system, or to ignore it, bears consequences for them in terms of experimental financial incentives. By so doing, we aim to make a key methodological contribution to HCI. Specifically, through the comparison of different experimental incentive strategies our results indicate that our translation of behavioural economics research methods to HCI can support the study of interactions with autonomous systems in the lab.

RELATED WORK

Our work builds upon prior research that has studied human interaction with autonomous systems involving both specialist and non-specialist users; approaches that influence the usage of autonomous systems; and the effect of displaying confidence information.

Specialist Applications of Autonomous Systems

Numerous studies have examined the effect of increased autonomy on users' performance with search tasks (e.g. finding victims in a disaster event) in a lab setting. For example, researchers have focused on the operation of robot teams by single or multiple users [17, 25, 36]. In more detail, participants either operated a team of manually operated robots or monitored a team of autonomously moving robots. Other work has investigated how different autonomy levels affect task allocation of multiple UAVs [27, 30]. More specifically, participants performed the search tasks either through manual or mixed-initiative task allocation of UAVs. Results from these studies showed that higher autonomy improved user accuracy with the tasks and reduced cognitive workload.

These studies focused on specialist applications, such as military [9, 19] or aviation [11, 29, 34], for which users would need and receive considerable amount of training. Moreover, these studies assumed that users would interact with autonomous systems. In contrast, our work focuses on non-specialist activities where users would not normally be trained to use the system and we examine whether the autonomous system would be used or not. The next subsection talks about autonomous systems in everyday life.

Autonomous Systems in Everyday Life

Researchers in HCI and UbiComp communities have explored the usage of emerging autonomous systems in the home environment. For example, Rodden et al. [31] used animated sketches to solicit views from people about current and future agent-based energy systems. Other studies instead investigated people's experience with existing smart products in the home, such as the Nest thermostat [37, 38] and Roomba [16, 35]. Evaluations of potential agent-based systems have also been conducted, such as for laundry management [6, 8] and tariff switching [1, 15]. Results from these studies suggest that users tend to be inclined to accept autonomous systems and integrate them in their day-to-day routines. For example, findings from the study by Costanza et al. [8] highlights that participants were able to integrate an agent-based system into their existing laundry practices.

However, results from other studies [37, 38] also revealed glitches in interaction with everyday autonomous systems. For example, after initial engagement with the Nest thermostats, its inability to match users' expectations led to frustrations [37]. Moreover, users became less engaged with the Nest thermostat over time, either overestimating or neglecting its capabilities [38]. As a result, people eventually missed opportunities where they could have saved energy and money. Complementing this work, and to address the issue of expectation mismatch, in this paper we present a study of whether displaying confidence can improve the utilisation of autonomous systems that help people in non-specialist activities. The next subsection talks about how users interact with autonomous systems when money is involved.

Financial Incentives with Autonomous Systems

In a series of studies by Dzindolet et al. [14], participants were asked to correctly identify whether a camouflaged soldier is present in an image or not. Additionally, a suggested answer from an automated aid would be shown after a participant has given an answer. In one of the studies, participants completed 200 trials and were paid $0.50 for each of 10 randomly selected trials, if their answer was correct. Participants were free to either choose their initial answers or the suggestions of the automated aid. Results show that more than 80% of the participants preferred their own answers over the automated aid. Furthermore, in a study by Alan et al. [1], participants were prepared to hand over tariff selection to an autonomous agent, even when the agent performance had financial consequences for them, but they were always keen to monitor the agent's actions at all times.

These studies highlight that users tend not to rely much on autonomous systems when there is an associated cost to reliance. For this reason, our study examined how the usage of autonomous systems can be improved by showing confidence information. In particular, we focused on whether displaying confidence can increase the usage of autonomous systems (i.e. by either reviewing its completed task or accepting its completed task without reviewing it), even though there is a risk of losing money.

Displaying Confidence Information

Displaying confidence information has mostly been researched with an aim to finding out how users interact with context-aware systems. Lemenson et al. [24] compared different

visualisation methods of confidence information for location-based services, while Antifakos et al. [2] asked participants to report whether they would check the settings automatically set up by a context-aware system given that they were in certain scenarios with different criticalities (e.g. while eating at a restaurant or while driving). Their findings suggest that participants were more willing to review the settings given that they were shown a display of confidence information, especially when the system's confidence level was low. Moreover, findings from a study by Lim et al. [26] suggest that displaying the confidence information of context-aware systems can affect users' understanding and impression of such systems in a variety of ways. A user's understanding and impression of a system can be improved when it has mostly high confidence levels. However, displaying confidence information can be harmful in situations where the system has mostly low confidence levels, as users tend to lose trust in its capabilities. In contrast to these studies, our work uses a functioning prototype to observe how users interact with autonomous agents rather than results elicited through reports of subjective preferences.

Antifakos et al. [3] examined whether displaying confidence information can improve the usage of context-aware memory aids. Their results suggest that users do perform better with the display of confidence, especially when the confidence level is high. Similarly, the findings of Dearman et al. [10] suggest that displaying confidence information can improve user performance in a search task using a location-based service application. In contrast to both studies, Rukzio et al. [32] found that displaying the confidence information of an automatic form filler slowed down users and caused them to make more errors as they often double-checked fields with lower confidence levels. Instead of focusing solely on performance, our work studies how confidence display can affect the usage of autonomous systems, especially when such usage has financial implications to the users.

Prior work [5, 18] has also investigated the effect of displaying the confidence information of self-driving cars. Results from these studies showed that displaying the confidence information reduced the time it took for drivers to take control of a self-driving car and allowed drivers to spend more time not looking at the road. In a study by McGuirl et al. [28], pilots were asked to complete a series of simulated flight exercises, requiring them to operate a number of manual tasks and monitor an automated system that would require users to take control at times. Results of this study indicate that pilots shown constantly updating confidence information were able to complete the flight tasks without failures and were better at estimating the accuracy of the automated system than pilots with only information about the overall reliability of the automated system (i.e. the accuracy of the automated system). In all three studies, the participants' choice to use the autonomous systems or not had no tangible consequence on them, e.g. it was not linked to any loss or gain of financial reward. In contrast, our work uses performance-based incentives for higher ecological validity.

Closer to our work, Desai et al. [12] reported a study investigating the effects of displaying the confidence of a moving robot that could move fully autonomously or in semi-autonomous mode (i.e. users can control the direction of its movement). Participants had to monitor a moving robot, help it pass obstacles and occasionally complete a secondary task (clicking a circle on the screen). In their study, participants were also rewarded based on task performance. The results showed that participants with the confidence information switched between full and semi-autonomy mode more than participants without the information. Particularly, participants were found switching to semi-autonomous mode whenever there is a drop in confidence, even though reliability did not change. This study required participants to actively monitor the autonomous robot, which enforced the interaction. In contrast, we focus on whether confidence information can increase the usage of autonomous systems in a scenario where they can choose to completely ignore it because in reality, people can choose to not use autonomous systems at all. In the next section, we detail our approach to the study method.

APPLYING BEHAVIOURAL ECONOMICS METHODS

Research concerning human interaction with non-specialist autonomous systems has adopted an in-the-wild approach, as a way to achieve realistic results [1, 6, 8, 37, 38]. While we agree with such an approach and believe that it is important to run field trials, we are *also* interested in studying interaction with autonomous systems in the lab. Lab studies allow for precise measurements to compare alternative experimental conditions, such as different interface features. Moreover, they tend to be faster and cheaper to run than field trials. However, studying the usage of autonomous systems in a lab setting involves the challenge of maintaining a high level of ecological validity.

In order to provide realism, we turned to experimental methods used in behavioural economics. Behavioural economics is concerned with the effects of psychological factors on people's economic decisions [21]. Typically, experiments by behavioural economists incorporate money in so called "choice situations", i.e. situations in which participants must choose from multiple options. For example, subjects would be asked whether to choose between an 85% chance to win $1000 (with a 15% chance to win nothing) and the alternative of receiving $800 for sure[4] [22].

More specifically our work was motivated by behavioural economics studies which involve the performance of repeated tasks with actual financial incentives: money is handed to participants based on their actions in the study. For example to investigate the effect of the perceived meaning of tasks on people's motivation to work, Ariely et al. [4] designed a study where participants were paid to complete a simple task: assembling a Lego model. Participants had the option to repeatedly complete this very same task several times, but each time at a reduced wage rate (first $3.00, then $2.70, then $2.40 and so on) [5]. Ours is not the first HCI project to turn to behavioural economics for inspiration. Previously, HCI researchers have

[4]Even though the first choice has the higher potential gain, most participants would prefer the guaranteed choice. This was posed as a hypothetical question, no money was handed to participants.

[5]Their results show that manipulating the task meaning can induce people to work for a significant lower pay rate.

suggested employing persuasion techniques based on effects studied by behavioural economists to promote healthy snack eating [23]. However, our approach is different and novel in that we turn to behavioural economics for experimental methods.

In our study, participants have the choice to perform one of two tasks: completing a *manual task* or checking the output produced by an autonomous system – we refer to this as the *agent task*. In such context, if participants perform the agent task, they give up the option to perform manual tasks, with an associated opportunity cost, because they have a limited amount of time and tasks allowed in the study. Our aim is to mimic a real-world situation whereby if a user chooses to invest time interacting with an autonomous system, doing so would cost the user time and effort. In the next section, we present our user study.

USER STUDY

A user study was designed and conducted to test the effectiveness of the confidence information of an autonomous software system. Specifically, we wanted to examine whether the confidence information affects users' decision to interact with an autonomous system helping users in an activity that can be considered mundane or common to various people (e.g. students, office workers, researchers). So we looked for an example autonomous software system, a so called 'agent', around which we could set up a credible scenario to play out in a lab study, and which could be related to tasks which would be natural for a population of university students and administrative staff. We settled for tasks and agents related to common *textual document* activities, such as typing up handwritten notes and proofreading text. Furthermore, we chose these tasks because we had access to the truth in each instance. This allowed us to automatically check the correctness of each submission and provide immediate feedback to participants about their performance. In the following subsections we first describe the autonomous agent and then detail the tasks used in the study.

OCR Agent

We designed an agent that automatically recognised handwritten text and converts into typed text – essentially an Optical Character Recognition (OCR) system. OCR applications are widespread and likely to be familiar (at least conceptually) to most of our participants.

The agent processes one document at a time, taking roughly 30 seconds. Participants were told that the agent may make mistakes, which would need correcting. These mistakes were incorrect type outs of characters that may look similar to other characters (e.g. the letters *n* and *h*). The OCR processing goes on in the background, autonomously. When the agent completes the task, a sound goes off indicating the availability of the results to the user (similar e.g., to receiving an incoming email message). Furthermore, as the agent completes tasks, they get added to a queue regardless of whether the user attended or ignored the previously completed task(s), somewhat similar to an email inbox.

Confidence	very low	low	med	high	very high
Mean	2.3	2	0.5	0	0.16
S.D.	1.5	1.4	0.55	0	0.4

Table 1. Character errors per confidence level. Note that because of randomness the *high* confidence level has fewer errors than *very high*.

Because of the possible mistakes made by the agent, users are required to 'review' or 'accept' the completed agent tasks. More specifically, for any completed agent task available in the queue, users had three options about how to deal with it:

1. *Review* – participants can view the task result to check and correct any errors.
2. *Blindly accept* – participants can blindly accept the agent result without reviewing it, essentially *fully accepting* the agent automation.
3. *Ignore* – participants can also opt not to review the completed agent task and just leave it in the queue. These tasks can be reconsidered at any later moment.

To ensure that the performance of the agent was consistent, as this could affect how participants use it, we adopted a Wizard-of-Oz approach. The agent was actually artificial, in that all the handwritten documents had been originally typed in by a researcher and errors were introduced in a controlled manner, to simulate 5 different levels of confidence on various documents. In particular, the five levels of confidence are: *very low, low, medium, high* and *very high*. Errors across different confidence levels were randomly distributed such that documents with *high* confidence level or higher had less character errors on average than documents with *medium* or lower confidence level. However, the confidence levels only roughly correlated to the number of documents with errors. In addition, the number of character errors in each document was varied in each confidence level (see Table 1 for error rates).

Alternative Tasks

To implement a choice situation where interacting with the agent would create an opportunity cost, we defined another task in addition to the **agent task** described above. This task, which we call the **manual task**, involves correcting the grammar of a 6-line long paragraph typed in English, checking for singular or plural agreement (*is, are, has, have*) and also for commonly mistaken possessive terms (*their* as *they're* and *its* as *it's*).

Although the agent helped users complete the transcription task, we wanted to make the work of reviewing agent tasks require more effort than completing the manual task. This was because in real-world situations, monitoring work completed by autonomous systems would require users to invest time that could otherwise be spent on doing other activities, especially at the beginning, when they have little experience with the system. For this reason, the documents processed in the agent task were written in a foreign language which would not be familiar to our study participants: Filipino[6]. This is to ensure that if users were to review the task, they would actually be comparing the handwritten and typed text. Such may not

[6]Anyone familiar with this language was excluded from our sample.

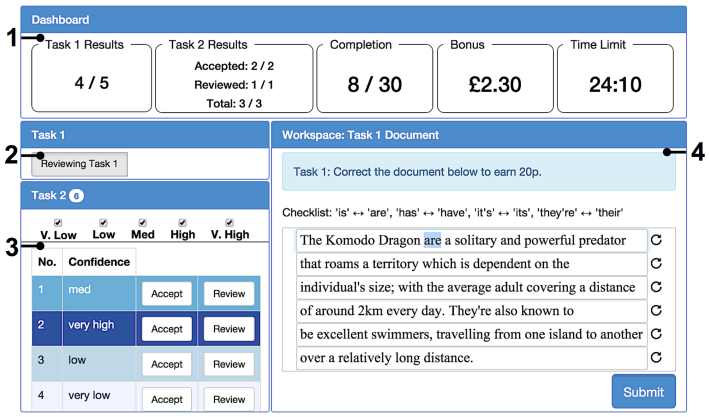

Figure 1. Screenshot of the interface, showing the dashboard (1), manual task switch (2), notification panel (3) and workspace (4). An example of a manual task is shown in the workspace. In this example, the highlighted word *are* must be replaced by the word *is*.

be the case if the manuscripts were in a common language (e.g. English or Spanish), where users may simply check the spelling of the typed text. In the next section, we detail the interactive system used in our lab study.

User Interface

We designed and developed an interactive system which simulated the scenario explained in the previous sections. Figure 1 shows the interface, which is divided into four main panes:

Dashboard (1). The dashboard contains statistical information about a user's status during the study. It displays the number of correct and submitted manual and agent tasks. Furthermore, the dashboard also shows the current reward and time limit.

Manual task switch (2). Allows users to switch to the manual task.

Notification panel (3). This panel shows agent tasks as they become available, where each row corresponds to one agent task. The *Review* button allows users to view the agent task, which will be shown in the current workspace. The *Accept* button allows users to blindly accept agent tasks. Each row also contains information about the confidence of the agent for that task. The rows are coloured according to the associated confidence (the higher the intensity, the higher the confidence). Furthermore, a filter function is available to help users filter the tasks based on the different confidence levels.

Workspace (4). The workspace shows the current task being performed. For example Figure 1 is showing the manual task, whereas Figure 2 shows an agent task with very low confidence.

Design

A 2×3 between-subjects study design was employed[7]. The *confidence information* was manipulated as an independent variable (IV), through the following conditions:

- Confidence – participants were able to see the agent's perceived confidence for each of its completed task.
- No-confidence – the confidence information was omitted. In addition, the agent tasks in the notification panel were not coloured.

We also manipulated the *incentive scheme* as an IV to validate the method we used in the study, with 3 conditions:

- No-incentive – participants were paid £6 for their participation, regardless of their performance in the study.
- Agent-incentive – participants were paid 50p for submitting an agent task without any mistakes and 20p for correcting all the grammatical mistakes in a manual task.
- Manual-incentive – participants were paid 20p for submitting an agent task without any mistakes and 50p for correcting all the grammatical mistakes in a manual task.

In the *agent-incentive* condition, the choice of payment reflects the amount of effort and time required to complete each of the tasks. Pilot studies revealed that manual tasks were completed in around 20 seconds in average, whereas the agent task took around 50 seconds. In short, the agent task took 50/20 = 2.5 times more time (and therefore effort) as the manual task. For the *manual-incentive* condition, we reversed the incentives used in the *agent-incentive* condition. This was done to double-check whether the level of incentives used in the *agent-incentive* were sufficient to motivate participants to choose one task more than the other. Furthermore, the *manual-incentive* condition was designed to negate or reduce the impact of factors other than the monetary reward that would affect users choosing the agent task. The next section details our hypotheses.

Hypotheses

We are particularly interested in how the confidence information affects participants' inclination to review or blindly accept agent tasks. The confidence information should make it possible for participants to know which agent tasks require lower effort (the ones with higher confidence). So we hypothesised that:

H1a – When confidence information is displayed, participants will use the agent **more**. In particular, they will complete (i.e. review or accept) a higher number of agent tasks than when the confidence information is omitted.

[7] A within-subject design was not possible because of the learning effect associated with the confidence information and also the types of errors in both tasks.

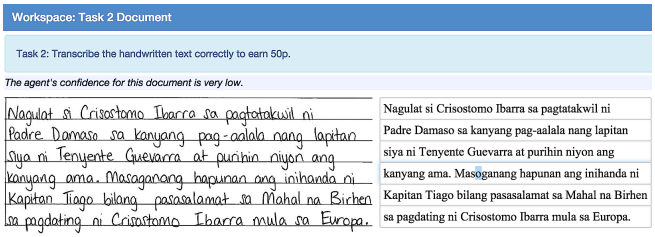

Figure 2. An example of an agent task. In this example, the highlighted letter *o* (4th line on the right side) must be replaced by the letter *a*.

H1b – Participants will complete more agent tasks with high confidence than agent tasks with lower confidence levels.

Secondly, confidence information should also inform users when the agent can be relied upon and when users need to intervene:

H2a – When confidence information is displayed, participants will rely more on the agent – i.e. they will blindly accept more agent tasks than when the confidence information is omitted.

H2b – Participants will accept more agent tasks with high confidence than agent tasks with lower confidence levels.

Additionally, we expect that our experimental method would affect the decision of users in choosing between completing the manual and the agent task. In particular, the different financial incentives imposed should influence users about which of the two tasks they should complete more. If so, this would validate our method. Our final hypothesis therefore is:

H3 – Participants in the *agent-incentive* condition will complete more agent tasks than manual tasks. Moreover, participants in the *no-incentive* and *manual-incentive* conditions will complete more manual tasks than agent tasks.

Participants

A total of 60 participants (39 female, 21 male) took part in the study, 10 per condition and 59 of these were members of the university: PhD, Masters and undergraduate students from a variety of disciplines (including Engineering, Languages, Business and Management, Law, Health and Social Sciences, and Geography). One participant works for the local council in data management for schools. The ages of these participants ranged from 18 to 43 years old ($M = 23.20$, $SD = 5.43$). As discussed above, the participants we recruited are educated to above average levels, but the tasks defined in our study are suitable for them.

Method

At the beginning of each experiment, participants were randomly assigned to one of the experimental conditions. Our participants were asked to complete up to 30 tasks in total within 30 minutes as accurately as they could. Crucially, participants were given the freedom to select whichever type of task they want to complete and were free to switch from one task to another at any given point in time. Indeed their selection of tasks was a key measure to quantify their inclination to use the autonomous agent. Participants paid based on performance (*agent-incentive* and *manual-incentive* conditions) were told that there is a limit of £10 to earn. Furthermore, participants in the *confidence* conditions were told that agent tasks have associated confidence levels. Details about how the confidence information was formed were not revealed to the participants. After these instructions, participants completed a 5-minute training period to help them gain familiarity with the system before starting the actual trial. Participants were shown how to switch between the two tasks during this training period. This is to ensure that they would not misunderstand how to complete the study, such as thinking that they need to complete all Task 1 documents first before doing Task 2 documents[8].

Data Collection

Data was collected through a combination of quantitative and qualitative techniques. The system automatically measured the following dependent variables:

Agent tasks completed – the proportion of agent tasks completed out of all completed tasks (the combination of reviewed and blindly accepted);

Agent tasks blindly accepted – the proportion of completed agent tasks that were not reviewed by the users out of all completed tasks;

Time – the average time taken (in seconds) for participants to complete the tasks, which can be interpreted as the amount of effort spent by participants;

Reward – the final reward received (in £) for participants in the *agent-incentive* and *manual-incentive* conditions;

Correct submissions – the proportion of tasks completed correctly out of all completed tasks;

Completed agent tasks per confidence level – the proportion of completed agent tasks by the users for each confidence level out of all completed agent tasks;

Blindly accepted agent tasks per confidence level – the proportion of blindly accepted agent tasks for each confidence level out of all blindly accepted agent tasks.

Moreover, participants were observed by a researcher throughout the study and interviewed at the end, to clarify their actions during the sessions. Each interview lasted approximately five minutes and was audio-recorded. Interviews were later coded through open codes for each experimental condition, then grouped in categories altogether through thematic analysis [7]. Open coding was completed per condition to identify main themes within each condition. Then, axial coding was completed for open codes across all conditions, to find the main themes for the whole study.

RESULTS

Quantitative Analysis

A total of 1088 manual tasks were completed and 768 of those were correct (70.59%). For agent tasks, 621 were completed

[8]The interviews confirmed that participants understood that it was possible to switch between the two tasks.

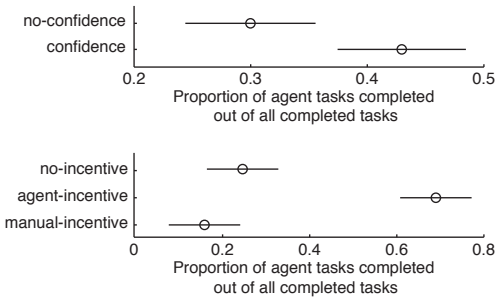

Figure 3. Means comparison for agent tasks completed across different displays of confidence information (top) and incentive schemes (bottom), with the 95% confidence bars (Tukey-HSD).

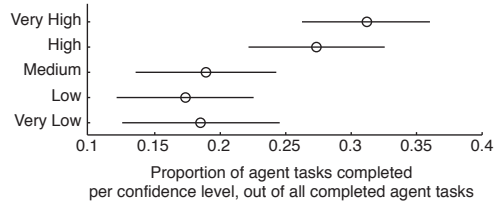

Figure 4. Means comparison for completed agent tasks per confidence level across all confidence levels, with the 95% confidence bars.

with 503 correct (81.00%). Furthermore, 126 of the completed agent tasks were blindly accepted (20.29%) and 99 of those blindly accepted tasks were correct (78.57%). In more detail, there were 50 completed agent tasks in the *no-incentive, no-confidence* condition, 94 in the *no-incentive, confidence*, 184 in the *agent-incentive, no-confidence*, 194 in the *agent-incentive, confidence*, 28 in the *manual-incentive, no-confidence* and 66 in the *manual-incentive, confidence*.

Proportion of agent tasks completed. A two-way ANOVA revealed a significant effect of both confidence information ($p < 0.05$) and incentive scheme ($p < 0.001$) on the proportion of agent tasks completed by participants. There was also no interaction effect. When confidence information was displayed, participants completed a higher proportion of agent tasks. A post-hoc Tukey test on the incentive schemes revealed that a higher proportion of agent tasks were completed in the *agent-incentive* condition ($M = 0.69$, $SD = 0.24$) than the *no-incentive* ($M = 0.25$, $SD = 0.25$) and *manual-incentive* ($M = 0.16$, $SD = 0.18$) conditions. Figure 3 shows the means comparison of the proportion of agent tasks completed, with 95% confidence intervals (Tukey-HSD), for this analysis.

Proportion of agent tasks blindly accepted. A two-way ANOVA revealed a significant effect of confidence information ($p < 0.05$) on the proportion of agent tasks blindly accepted by participants, with a higher proportion of tasks being blindly accepted in the *confidence* condition. No statistically significant differences were found based on incentive schemes and there was also no interaction effect.

Proportion of agent tasks completed per confidence level. A one-way ANOVA revealed a significant effect of confidence level ($p < 0.001$). A post-hoc Tukey test revealed that there were significantly more completed agent tasks with *very high* confidence level ($M = 0.31$, $SD = 0.21$) than agent tasks with *medium* ($M = 0.19$, $SD = 0.07$), *low* ($M = 0.17$, $SD = 0.10$) and *very low* ($M = 0.19$, $SD = 0.07$) confidence level. Figure 4 shows the means comparison of the proportion of agent tasks completed per confidence level for all confidence levels.

Proportion of blindly accepted agent tasks per confidence level. A one-way ANOVA revealed a significant effect of confidence level ($p < 0.001$). A post-hoc Tukey test revealed that there were significantly more blindly accepted agent tasks with *very high* confidence level ($M = 0.55$, $SD = 0.27$) than agent tasks with *high* ($M = 0.30$, $SD = 0.18$), *medium* ($M = 0.05$, $SD = 0.09$), *low* ($M = 0.03$, $SD = 0.05$) and *very low* ($M = 0.07$, $SD = 0.26$) confidence level. The *high* confidence agent tasks were also blindly accepted significantly more than agent tasks with *medium*, *low* and *very low* confidence level. Figure 5 shows the means comparison of the proportion of blindly accepted agent tasks per confidence level for all confidence levels.

Reward. A two-way ANOVA revealed a significant effect of incentive scheme ($p < 0.05$), but revealed no statistical significance across both displays of confidence information, with no effect of interaction between the two. There were significantly more reward earned in the *manual-incentive* ($M = 9.52$, $SD = 0.80$) than in the *agent-incentive* condition ($M = 8.65$, $SD = 1.62$).

Time. Participants took longer to complete agent tasks ($M = 49.48$, $SD = 24.84$) than manual tasks ($M = 35.22$, $SD = 11.29$) and a one-way ANOVA test indicates that this difference is significant ($p < 0.001$). Furthermore, a two-way ANOVA revealed a significant effect of incentive scheme ($p < 0.05$), but revealed no statistical significance across both displays of confidence information, with no effect of interaction between the two. A post-hoc Tukey test revealed that participants in the *agent-incentive* condition took significantly more time ($M = 49.60$, $SD = 11.62$) than participants in both the *no-incentive* ($M = 37.03$, $SD = 12.91$) and *manual-incentive* ($M = 37.07$, $SD = 14.13$) conditions. Figure 6 shows the means comparison of average task time completion for all incentive schemes.

Correct submissions. A two-way ANOVA revealed no statistical significance across incentive schemes and displays of confidence information, with no effect of interaction between the two.

Summary. In summary, the quantitative analysis of our data revealed that the display of *confidence* led participants to work on a higher proportion of agent tasks (top of Figure 3) and also blindly accept a higher proportion of agent tasks. Within the *confidence* condition, participants were more likely to work on tasks with *very high* confidence than any other tasks (Figure 4) and to blindly accept tasks with *very high* confidence more than tasks with *high* confidence, and these in turn more than tasks with lower levels of confidence (Figure 5). In terms of reward, the *agent-incentive* condition led participants to work on a higher proportion of agent tasks (bottom of Figure 3). In the *manual-incentive* condition participants gained a higher reward, while in the *agent-incentive* condition they spent more time on average per task (Figure 6).

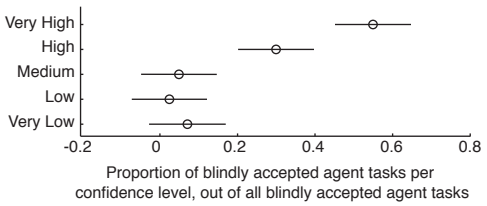

Figure 5. Means comparison for blindly accepted agent tasks per confidence level across all confidence levels, with the 95% confidence bars.

Qualitative Analysis

Through the interviews, participants gave us insight on the choices and strategies they adopted during the study.

Selecting which tasks to complete

Our participants indicated that a number of factors influenced their choice between completing agent or manual tasks, including how easy and how challenging the tasks were perceived to be, as well as the reward associated with the task.

Easier. All but one participants in each of the *no-incentive* (19) and *manual-incentive* (19) conditions, and few participants (3) in the *agent-incentive* condition reported that they preferred manual tasks because they perceived it was *easier* and *quicker* to complete them. This was sometimes related to the tasks being written in a familiar language. Similarly, four other participants (2 in the *agent-incentive* condition and 1 from each of the *no-incentive* and *manual-incentive* conditions) mentioned that agent tasks were more challenging. However they reported such challenge to be a reason to complete them. Conversely, half of the participants in the *agent-incentive* condition (note it was only in this condition) told us that they found agent tasks easier, and this was a factor for preferring them. These participants reported that it was easier for them to compare snippets of text rather than completing a task that required grammatical reasoning.

Money matters, or not. Most of the participants in the *manual-incentive* and *agent-incentive* conditions (14 in each) also mentioned that the reward was a contributing factor for choosing the better paid tasks. At the same time, 5 participants in the *manual-incentive* condition and 3 in the *agent-incentive* condition were dismissive about the reward being a factor. Other participants mentioned the uncertain reliability of the agent as a reason for not completing agent tasks at all, rather than the reward.

Switching between tasks. Overall 44 participants performed a combination of manual and agent tasks, while the remaining 16 performed only tasks of one type. Various factors were reported as reasons to switch type of tasks, including wanting to have a bit of variation, and curiosity to try both tasks.

The choice of what kind of tasks to complete was driven by various factors, with a certain degree of subjectivity. Ease of completion, challenge and financial reward level all played a role. In the next subsection, we consider how the confidence information affected users behaviour in the study.

Utilising the confidence information

Interpretation. The 30 participants in the *confidence* condition had various interpretations about what the confidence

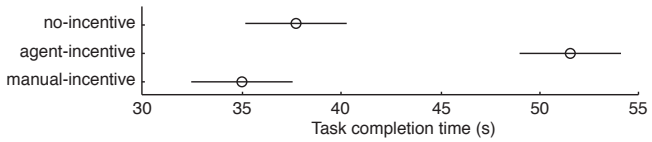

Figure 6. Means comparison for task completion time across different incentive schemes, with the 95% confidence bars.

levels meant. For 17 of them the tasks with *high* and *very high* confidence were very accurate, and only the tasks with lower confidence levels had mistakes. Another 7 participants felt that the confidence information was not an important indicator. The remaining 6 participants only completed manual tasks, and hence entirely ignored the confidence information.

Strategies around confidence. We noticed from our observations that the 17 participants who gave importance to confidence devised different tasks completion strategies leveraging this information. Such observation was confirmed through the interviews.

Prioritise high confidence agent tasks. The vast majority of participants, 15, prioritised tasks with *high* or *very high* confidence, before working on either lower confidence agent tasks or manual tasks. They perceived these tasks would likely have the least amount of errors and therefore would require the least amount of effort. Furthermore, 13 of these participants went as far as blindly accepting agent tasks with higher confidence because they felt that for these they could rely on the agent to do a good job.

Prioritise low confidence agent tasks. Conversely, 2 participants focused on low confidence agent tasks first, and left higher confidence ones for later. While all of these participants also believed that the high confidence agent tasks were reliable enough to not need reviewing, they chose to go for the lower confidence tasks because they wanted a more challenging task. These 2 participants also mentioned that they intended to blindly accept high confidence tasks at the end, when they would have had little time left. However, they did not manage to do so, because by that point they had already reached the limit of 30 tasks or £10. In contrast, the remaining 13 participants in the *confidence* condition completed their trials without reference to the confidence information. As a result, their approach to completing tasks was similar to the approach of participants in the *no-confidence* condition, which we will detail in the next subsection.

Making sense of the agent without confidence information

All the participants who engaged agent tasks in the *no-confidence* condition (19 in total) and 8 of those in the *confidence* condition reported reviewing (rather than blindly accepting) agent tasks because they could not rely on the agent. For example P1 (*no-incentive, no-confidence*), a 27-year-old female PhD student in Social Statistics and Demography, told us: *"I've written in foreign languages before using just the computer into Word. When it changes, it gives you suggestion to change the grammar [and] it's not usually correct, [especially] if you are using a foreign language"*.

Eight participants in the *no-confidence* condition blindly accepted agent tasks. Some reported doing so to "*gamble*" or

"*try out their luck*", in the hope to earn money easily through the study. Others pressed the *Accept* button when they were running out of time as an attempt to earn as much money as possible. There were also 3 participants in the *no-confidence* condition who blindly accepted agent tasks because they were bored and "*wanted to try the software*".

DISCUSSION

Financial incentives and experimental method

The statistical analysis of our results revealed that the incentive scheme had an effect on the type of tasks that participants chose to complete. A higher proportion of agent tasks were completed in the *agent-incentive* condition than both the *no-incentive* and *manual-incentive* conditions. In other words, participants were sensitive to the financial incentives, and completed more of the type of tasks for which they received higher incentives. This result confirms our hypothesis H3, and validates our method, in that it demonstrates that the use of financial incentives was successful in motivating participants to do a specific task. In the *no-incentive* condition (where participants received a fixed £6 reward regardless of performance and task choice), participants completed more of the manual tasks, which is the one that requires the least effort. Users' sensitivity to financial incentives in the *agent-incentive* condition (i.e., more agent tasks get done) also indicates that participants are more inclined to use the agent when it provides higher utility than the manual task. In our study, experimental financial reward mimicked a situation in which the agent performs a task that is practically useful to participants (a real life example would be saving money on the energy bills by automatically controlling the thermostat). However, it should be noted as a limitation that the game-like nature of our experiment (including its limited duration) may have influenced participants to give more importance to the financial incentives than would be observed in real life. In other words, participants in the study may feel compelled to try and "win" as much as they can, just because it is a game [13].

Even though our quantitative data clearly shows that the incentive scheme and the confidence information both had statistically significant effects on participants' behaviours, in the interviews participants suggested that a more complex and varied set of factors influenced the choice of tasks to complete. Most participants suggested that reward was only *one* contributing factor for preferring a task, while some went as far as completely dismissing the idea that the reward influenced their behaviour. Other reported factors included how easy or how challenging the task was perceived to be. At the same time, only participants in the *agent-incentive* condition described agent tasks as easier, and these are the tasks for which they received higher incentives. Furthermore, the general majority of participants reported manual tasks to be easier, and hence preferable. Therefore, the perception of a task as 'easy' seems to be influenced by the financial incentives. It is possible that such bias was unconscious, or that participants felt embarrassment to acknowledge that they are driven by money. Such contrast between the quantitative results and the findings from the interviews reminds us that self-report may not always be dependable on its own, especially when attitudes towards financial incentives are involved.

In addition, the incentive scheme also had a statistically significant effect on the financial reward gained. Participants in the *manual-incentive* condition (where the manual task was rewarded more) earned more money than participants in the *agent-incentive* condition, suggesting that the manual task was easier than the agent task, as we intended. Such difference in effort required was further confirmed by another result of our analysis: participants took longer to complete agent tasks, on average, than to complete manual tasks.

Displaying the confidence information

The confidence information made a difference in how our participants interacted with the agent. We specifically hypothesised that there would be more agent tasks completed in the *confidence* condition than in the *no-confidence* condition (H1a). Our statistical analysis shows that a higher proportion of agent tasks were performed when the system displayed the confidence information, confirming hypothesis H1a. In particular, participants completed more tasks with *very high* confidence level than tasks with lower levels of confidence, according to our hypothesis H1b. These results suggest that the different confidence information informed users about the amount of effort required before actually starting the tasks. Similarly, participants in the *confidence* condition blindly accepted a higher proportion of agent tasks, than in the *no-confidence* condition, confirming H2a. Furthermore, agent tasks with *very high* confidence were blindly accepted more than those with *high* confidence, and these in turn were blindly accepted more than tasks with lower confidence, confirming H2b.

The display of confidence information enabled participants to rely on the autonomous agent more. This result is in line with prior work on displaying confidence information [5, 18, 28]. In turn, and as expected, our participants were unable to make an informed decision about using the agent when they had no confidence information. This result is also similar to findings from prior studies [1, 38], even though our work is based on a different study method and different application (not energy related). To further support the quantitative data on this aspect, the interviews revealed a striking contrast between the *confidence* and *no-confidence* conditions. On the one hand, when confidence information was displayed most participants reported taking it into account for gauging their expectations about the performance of the agent, and in turn for choosing which tasks to perform. On the other hand, without confidence information available, participants described how they resorted to alternative ways to make sense of the agent, and to set their expectations. For example they referred to prior experience with systems that they considered similar, such as spell checking software. However, such similarities may be based on superficial aspects of the systems, and hence be insubstantial, with the associated risk of generating incorrect expectations. To summarise, displaying the agent's confidence information allowed users to form strategies about how to utilise the system based on their own attitude. Hence, the confidence information also increased the usage of the autonomous agent. We elaborate on these strategies in the next subsection.

Subjective perception and attitude

In general, our participants employed different strategies in utilising the confidence information, reflecting different personal attitudes toward autonomous systems. For example, some participants dismissed the confidence information, and the agent operation in general, based purely on their experience with other different computational systems. Other participants reported a preference for maintaining some form of control, similar to what has been reported in prior work [1]. Others still acknowledged the meaning of the confidence information, but they favoured manual tasks, or agent tasks with lower confidence because they considered them more challenging, and hence rewarding. This finding aligns with the results of a study by Ariely et al. [4], who found that participants completing meaningful tasks were more motivated to work than participants who are working on less meaningful tasks. In our study, earning rewards by reviewing agent tasks was recognised by some participants as a more meaningful endeavour than simply earning rewards by blindly accepting agent tasks.

The interview data also revealed different user perceptions of the confidence information. Most participants were able to pick up on how well the confidence levels correlated to the reliability of the agent's output. These participants would describe that "*the chances were a lot higher*" for agent tasks with a *high* confidence level or higher to be correct, while they felt that "*there probably would be at least one mistake*" for agent tasks with a *medium* confidence level or lower. However, not all participants perceived that the confidence information related to the agent's capability. Some participants felt that they "*couldn't really distinguish a pattern*" and that the agent was only "*saying its confidence, it's still not a 100% positive*". This mindset of not relying on the confidence information emerged from participants who reported that they do not trust systems that can be considered similar to the one used in this study. It should be noted that only a minority of the participants (13) in the confidence condition reported such an attitude. Indeed the quantitative results indicate that, *in general*, displaying the confidence information makes a significant difference to the usage of autonomous systems. In summary, the user's perception of the display of confidence information is affected by the user's willingness to trust the systems that produce it. In our study, even though the confidence information provided was a reliable estimation of the correctness of the agent's output, there were still participants who disregarded it – a result of their reservations about trusting autonomous systems.

Reflecting on overall performance

No statistically significant effects of confidence information were found on the total reward gained by participants, nor on task completion time. The reward can be considered a proxy for the participant's overall performance in the experiment. While this finding is not conclusive (a larger sample size may reveal statistically significant differences), it does suggest that the confidence information did not influence overall performance. This result is perhaps counter-intuitive, because confidence led participants to blindly accept a higher proportion of (higher confidence) tasks, making them in principle more productive. Indeed, this result is in contrast with previous studies showing that displaying confidence information can improve user performance [3, 28, 32]. One possible explanation here is that the time gained by blindly accepting tasks was spent in an unproductive way (unproductive in terms of the experiment financial reward). Indeed the interviews suggest that some participants preferred tasks that are more challenging, rather than easier, or tasks for which they have more control, because they were generally sceptical about the agents' abilities and disregarded the confidence information.

Limitations

The work presented in this paper relies on the availability of accurate confidence information, such as when the system uses an appropriate model to learn the data. However, it is important to point out that this may not always be the case. Further research is needed to evaluate the effects of unreliable confidence information. Furthermore, while our method places considerable emphasis on financial rewards as a motivational factor for using (or ignoring) the autonomous system, the interviews revealed that a variety of other factors are also at play (e.g. curiosity, challenge, etc.). While our method has proven to be flexible enough to allow these factors to emerge, future work should investigate situations where financial effects are not in the picture at all. Lastly, future studies should investigate longer term effects and also how people would react to finer- or coarser-grained confidence levels.

CONCLUSION

In this paper, we have presented a lab study with 60 participants, designed to investigate whether the display of confidence information influences users attitude towards autonomous systems, particularly those for non-specialist applications. A combination of quantitative and qualitative data revealed that when confidence information is available users are more likely to take advantage of the agent. This result can be explained through the observation that users can be guided in selecting which agent tasks to concentrate on by displaying the confidence information.

An important implication of our work, then, is that if at all possible confidence information should be included in the feedback from autonomous and smart systems to increase the chances of their uptake. Moreover, through a comparison of the effects of different incentive schemes our study also demonstrates that our participants were sensitive to different reward mechanisms. Such findings suggest that it is possible to design reward mechanisms and experimental tasks to realistically evaluate interactions with autonomous systems in a controlled lab setting. We hope that our work will motivate other researchers to take advantage of this method.

ACKNOWLEDGEMENTS

This research was funded, in part, by the EPSRC ORCHID project (EP/I011587/1). Data URL: http://doi.org/bbnz. Study approved by U.Southampton FPSE Ethics Committee (ref: 14292).

REFERENCES

1. Alper Alan, Enrico Costanza, Joel Fischer, Sarvapali D. Ramchurn, Tom Rodden, and Nicholas R. Jennings. 2014. A Field Study of Human-agent Interaction for Electricity

Tariff Switching. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems (AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 965–972. http://dl.acm.org/citation.cfm?id=2615731.2617400

2. Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. Towards Improving Trust in Context-aware Systems by Displaying System Confidence. In *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices &Amp; Services (MobileHCI '05)*. ACM, New York, NY, USA, 9–14. DOI: http://dx.doi.org/10.1145/1085777.1085780

3. Stavros Antifakos, Adrian Schwaninger, and Bernt Schiele. 2004. Evaluating the Effects of Displaying Uncertainty in Context-Aware Applications. In *UbiComp 2004: Ubiquitous Computing*, Nigel Davies, ElizabethD. Mynatt, and Itiro Siio (Eds.). Lecture Notes in Computer Science, Vol. 3205. Springer Berlin Heidelberg, 54–69. DOI:http://dx.doi.org/10.1007/978-3-540-30119-6_4

4. Dan Ariely, Emir Kamenica, and Draen Prelec. 2008. Man's search for meaning: The case of Legos. *Journal of Economic Behavior & Organization* 67, 34 (2008), 671 – 677. DOI:http://dx.doi.org/10.1016/j.jebo.2008.01.004

5. Johannes Beller, Matthias Heesen, and Mark Vollrath. 2013. Improving the Driver-Automation Interaction: An Approach Using Automation Uncertainty. *Human Factors* 55, 6 (2013), 1130–1141. DOI: http://dx.doi.org/10.1177/0018720813482327

6. Jacky Bourgeois, Janet van der Linden, Gerd Kortuem, Blaine A. Price, and Christopher Rimmer. 2014. Conversations with My Washing Machine: An In-the-wild Study of Demand Shifting with Self-generated Energy. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 459–470. DOI: http://dx.doi.org/10.1145/2632048.2632106

7. Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

8. Enrico Costanza, Joel E. Fischer, James A. Colley, Tom Rodden, Sarvapali D. Ramchurn, and Nicholas R. Jennings. 2014. Doing the Laundry with Agents: A Field Trial of a Future Smart Energy System in the Home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 813–822. DOI: http://dx.doi.org/10.1145/2556288.2557167

9. E Greef de Tjerk, FR Arciszewski Henryk, and Mark A Neerincx. 2010. Adaptive automation based on an object-oriented task model: Implementation and evaluation in a realistic c2 environment. *Journal of Cognitive Engineering and Decision Making* 4, 2 (2010), 152–182.

10. David Dearman, Alex Varshavsky, Eyal De Lara, and Khai N. Truong. 2007. An Exploration of Location Error Estimation. In *Proceedings of the 9th International Conference on Ubiquitous Computing (UbiComp '07)*. Springer-Verlag, Berlin, Heidelberg, 181–198. http://dl.acm.org/citation.cfm?id=1771592.1771603

11. Frederic Dehais, Vsevolod Peysakhovich, Sébastien Scannella, Jennifer Fongue, and Thibault Gateau. 2015. "Automation Surprise" in Aviation: Real-Time Solutions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2525–2534. DOI: http://dx.doi.org/10.1145/2702123.2702521

12. Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of Robot Failures and Feedback on Real-time Trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction (HRI '13)*. IEEE Press, Piscataway, NJ, USA, 251–258. http://dl.acm.org/citation.cfm?id=2447556.2447663

13. Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments (MindTrek '11)*. ACM, New York, NY, USA, 9–15. DOI: http://dx.doi.org/10.1145/2181037.2181040

14. Mary T. Dzindolet, Linda G. Pierce, Hall P. Beck, and Lloyd A. Dawe. 2002. The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors* 44, 1 (2002), 79–94. DOI: http://dx.doi.org/10.1518/0018720024494856

15. Joel E. Fischer, Sarvapali D. Ramchurn, Michael Osborne, Oliver Parson, Trung Dong Huynh, Muddasser Alam, Nadia Pantidi, Stuart Moran, Khaled Bachour, Steve Reece, Enrico Costanza, Tom Rodden, and Nicholas R. Jennings. 2013. Recommending Energy Tariffs and Load Shifting Based on Smart Household Usage Profiling. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI '13)*. ACM, New York, NY, USA, 383–394. DOI: http://dx.doi.org/10.1145/2449396.2449446

16. Jodi Forlizzi and Carl DiSalvo. 2006. Service Robots in the Domestic Environment: A Study of the Roomba Vacuum in the Home. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction (HRI '06)*. ACM, New York, NY, USA, 258–265. DOI: http://dx.doi.org/10.1145/1121241.1121286

17. Michael A. Goodrich, Timothy W. McLain, Jeffrey D. Anderson, Jisang Sun, and Jacob W. Crandall. 2007. Managing Autonomy in Robot Teams: Observations from Four Experiments. In *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction (HRI '07)*. ACM, New York, NY, USA, 25–32. DOI: http://dx.doi.org/10.1145/1228716.1228721

18. Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 2013. Presenting System Uncertainty in Automotive UIs for Supporting Trust Calibration in Autonomous Driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '13)*. ACM, New York, NY, USA, 210–217. DOI: http://dx.doi.org/10.1145/2516540.2516554

19. Tove Helldin, Ulrika Ohlander, Göran Falkman, and Maria Riveiro. 2014. Transparency of Automated Combat Classification. In *Engineering Psychology and Cognitive Ergonomics*. Springer, 22–33. DOI: http://dx.doi.org/10.1007/978-3-319-07515-0_3

20. N. R. Jennings, L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden, and A. Rogers. 2014. Human-agent Collectives. *Commun. ACM* 57, 12 (Nov. 2014), 80–88. DOI:http://dx.doi.org/10.1145/2629559

21. Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

22. Daniel Kahneman and Amos Tversky. 1984. Choices, values, and frames. *American Psychologist* 39, 4 (1 April 1984), 341–350. DOI: http://dx.doi.org/10.1037/0003-066x.39.4.341

23. Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011. Mining Behavioral Economics to Design Persuasive Technology for Healthy Choices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 325–334. DOI:http://dx.doi.org/10.1145/1978942.1978989

24. Hendrik Lemelson, Thomas King, and Wolfgang Effelsberg. 2008. A Study on User Acceptance of Error Visualization Techniques. In *Proceedings of the 5th Annual International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services (Mobiquitous '08)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, Article 53, 6 pages. DOI: http://dx.doi.org/10.4108/ICST.MOBIQUITOUS2008.3889

25. M. Lewis, Huadong Wang, Shih-Yi Chien, P. Scerri, P. Velagapudi, K. Sycara, and B. Kane. 2010. Teams organization and performance in multi-human/multi-robot teams. In *Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on*. 1617–1623. DOI: http://dx.doi.org/10.1109/ICSMC.2010.5642379

26. Brian Y. Lim and Anind K. Dey. 2011. Investigating Intelligibility for Uncertain Context-aware Applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. ACM, New York, NY, USA, 415–424. DOI: http://dx.doi.org/10.1145/2030112.2030168

27. Lanny Lin and Michael A. Goodrich. 2015. Sliding Autonomy for UAV Path-Planning: Adding New Dimensions to Autonomy Management. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '15)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1615–1624. http://dl.acm.org/citation.cfm?id=2772879.2773357

28. John M. McGuirl and Nadine B. Sarter. 2006. Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48, 4 (2006), 656–665. DOI: http://dx.doi.org/10.1518/001872006779166334

29. Raja Parasuraman, Robert Molloy, and Indramani L. Singh. 1993. Performance Consequences of Automation-Induced 'Complacency'. *The International Journal of Aviation Psychology* 3, 1 (1993), 1–23. DOI: http://dx.doi.org/10.1207/s15327108ijap0301_1

30. Sarvapali Ramchurn, Joel Fischer, Yuki Ikuno, Feng Wu, Jack Flann, and Antony Waldock. 2015. A Study of Human-Agent Collaboration for Multi-UAV Task Allocation in Dynamic Environments. (2015). http://www.orchid.ac.uk/eprints/id/eprint/242

31. Tom A. Rodden, Joel E. Fischer, Nadia Pantidi, Khaled Bachour, and Stuart Moran. 2013. At Home with Agents: Exploring Attitudes Towards Future Smart Energy Infrastructures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1173–1182. DOI: http://dx.doi.org/10.1145/2470654.2466152

32. Enrico Rukzio, John Hamard, Chie Noda, and Alexander De Luca. 2006. Visualization of Uncertainty in Context Aware Mobile Applications. In *Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '06)*. ACM, New York, NY, USA, 247–250. DOI: http://dx.doi.org/10.1145/1152215.1152267

33. Thomas B Sheridan and William L Verplank. 1978. *Human and computer control of undersea teleoperators*. Technical Report. DTIC Document.

34. P.J. Smith, C.E. McCoy, and C. Layton. 1997. Brittleness in the design of cooperative problem-solving systems: the effects on user performance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 27, 3 (May 1997), 360–371. DOI: http://dx.doi.org/10.1109/3468.568744

35. Ja-Young Sung, Lan Guo, RebeccaE. Grinter, and HenrikI. Christensen. 2007. My Roomba Is Rambo: Intimate Home Appliances. In *UbiComp 2007: Ubiquitous Computing*, John Krumm, GregoryD. Abowd, Aruna Seneviratne, and Thomas Strang (Eds.). Lecture Notes in Computer Science, Vol. 4717. Springer Berlin Heidelberg, 145–162. DOI: http://dx.doi.org/10.1007/978-3-540-74853-3_9

36. Jijun Wang, Michael Lewis, and Paul Scerri. 2006. Cooperating robots for search and rescue. In *Agent Technology for Disaster Management Workshop at AAMAS*, Vol. 6.

37. Rayoung Yang and Mark W. Newman. 2013. Learning from a Learning Thermostat: Lessons for Intelligent Systems for the Home. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 93–102. DOI: http://dx.doi.org/10.1145/2493432.2493489

38. Rayoung Yang, Mark W. Newman, and Jodi Forlizzi. 2014. Making Sustainability Sustainable: Challenges in the Design of Eco-interaction Technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 823–832. DOI: http://dx.doi.org/10.1145/2556288.2557380