UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL AND APPLIED SCIENCES

Electronics and Computer Science

Visual cues: Changing how people perceive smart systems' performance

by

Pedro García García

Thesis for the degree of Doctor of Philosophy

August 2017

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL AND APPLIED SCIENCES
Electronics and Computer Science

Doctor of Philosophy

VISUAL CUES: CHANGING HOW PEOPLE PERCEIVE SMART SYSTEMS' PERFORMANCE

by Pedro García García

In this thesis, we report twelve studies. In more detail, four lab studies and eight follow-up studies on the crowd-sourcing platform designed to investigate the potential of *visual cues* to influence users' perception of three smart systems: a vacuum robot, a handwriting recognition and a part-of-speech tagging system. The findings from the first three studies indicate that *physical motion cues* can influence people's perception of vacuum robots' performance. The subsequent three studies indicate that indeed *animation cues* can influence a participant's perception of handwriting recognition and part-of-speech tagging systems' performance. The subsequent three studies, designed to try and identify an explanation of this effect, suggest that it is related to the participants' mental model of the smart system. The last three studies were designed to characterise the effect more in detail, and they revealed that different detail of animation does not seem to create substantial differences and that the effect persists even when the system's performance decreases, but only when the difference in performance level between the systems being compared is small. Finally, the last study focused on analysing the effect of varying the speed of the animation, and we found that the effect persists even the variation of speed in the animation.

Contents

List of Figures

Declaration of Authorship

I, Pedro García García , declare that the thesis entitled *Visual cues: Changing how people perceive smart systems' performance* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as: (Garcia et al., 2016)

Signed:.....

Date:.....

Acknowledgements

This work has been possible with the guidance, ideas and support of my supervisor Enrico and Gopal. You guys help me to reach new horizons that I did not expect to see before I started my PhD. Also, I want to recognise the help, ideas, chats, time, and friendship of my colleges Elham, Diana, Jhim, Vangelis, Mike, and Jacob, I can not imagine having done my research without all of you. Thanks for everything you did in the office and outside for me.

"The greatest gift of life is friendship, and I have received it." (Hubert H. Humphrey). I want to say thanks to all my friends that become family in this journey. Enrique Cuan (Kike) and Eduardo Pérez (Lalo), I do not have words to express all the gratitude for your friendship and support. I will always remember all the talks, poker nights, and food we cooked in those days that we felt disheartened. My ladies that never left me behind, Aurea, Ale Vergara, Ana Aranda, Rubí, Pamela, and Rosie. I will keep in my heart the trips, talks, walks, coffees, and all the tears we share in those hard moments. Moreover, I want to acknowledge my friends Luna, Miguel, Pamela and Brenda that are more than just friends, they are my siblings. You always motivate me in every moment. Luna, I do not have words to say how much I love you, you are always there to listen to me and make me laugh. Miguel (my bro) and Pamela, thanks for your friendship and always cheer me up does not matter what. Brenda, you are an incredible woman, and I am so happy that you are part of my family (my little sister). Thanks for all the dinners, cinema nights, talks, and taking care of Tona.

In Southampton, I found my second home in Pansy Road and especially my new family. Weronika, Paulina, Eli, Jhim, Paul, and Harry, you made my days in that amazing house. Thanks for all the nights, talks, support, and details you had with me. I will remember my birthdays and nights we spent talking until the sunrise.

Finally, I want to acknowledge those who always are there for me without hesitating, my family. Mum and Dad, you are the reason of my existence, and you always support me even I have a strong character, thanks for all the calls and hear me when I need it more. I love you too much and always it is a glory to say that you are my parents. This accomplishment is yours as well. My siblings that always are with me, Carlos you help me and make sure that I am fine even I am a grown man. However, you always see me like your little boy, thanks for that. Tona, thanks for hearing me and be my best friend. I know that we fight too much and we get angry. However, you are the only one who I will give everything without thinking twice, thanks for arriving at my life. Finally, my niece who always play with me even though I was so far, you never left to talk with me on the weekends and Cony for giving me her. I appreciate to my aunts, uncles, and grandpa for all their love and support.

“Para finalizar quiero agradecer a las personas que sin importar nada siempre estn y estuvieron ahí, mi familia. Papás ustedes son mi razón de ser y siempre me han apoyado a pesar de mi carácter, gracias por todos esos días de estar platicando conmigo por webcam y siempre escucharme. Los amo mucho y es un orgullo poder decir que ustedes son mis padres. Este logro es también de ustedes. A mis hermanos que siempre están conmigo, Carlos siempre estás apoyando y procurando que esté bien a pesar de que ya estoy viejo, pero aún así me sigues tratando como tu hermanito, gracias por ello. Tona, gracias por escucharme y ser mi mejor amigo, se que peleamos mucho y nos enojamos, pero tu eres esa persona por la que daría todo, gracias por llegar a mi vida. Finalmente, a mi sobrina Michelle, por jugar conmigo a pesar de estar lejos, nunca dejaste de marcarme los fines para platicar un rato y a Cony por regalarme mi sobrina. Agradezco a toda mi familia por el cario y apoyo que me dan.”

Special acknowledgement to my friend, confident, and girlfriend Mariana. Sometimes the path is heavy. However, only the right person can help you to reach your goal. Thanks for all the calls and crazy moments we share.

I want to recognise all the work and support of my sponsor CONACYT and SICYT. Thanks for the opportunity you gave me to reach my goal.

This thesis is in memory of my angel and mum (grandma) Altagracia Reynoso Peralta, thanks for keeping me safe and all the love that only a mum can give.

"Esta tesis es en memoria a mi ángel y mamá Altagracia Reynoso Peralta, gracias por cuidarme y por el amor que solo una mamá puede dar. "

Nomenclature

HCI	Human Computer Interaction
HWR	**H**and**W**riting **R**ecognition
IoT	Internet of Things
MTurk	Amazon **M**echanical **Turk**
NHL	non-human-like
POS	part-of-speech

Chapter 1

Introduction

There is a growing number of smart systems that can automatically perform actions on behalf of users. Such systems are becoming increasingly widespread for non-specialist applications, such as in school environments, where the aim is to support teachers in their task of teaching and managing their classroom activities. Examples of this particular type of system include: Learning Management System (Cheema et al. (2016)), smart boards[1] (the smart board uses touch detection for user input in the same way as PC input devices allowing users to interact with presentations that are displayed in whiteboards), and smart tables[2] (the smart table can track and identify multiple objects simultaneously when placed on top of its surface and user can interact with the data and information that are displayed on it). Additionally, more systems are becoming part of the domestic environment to help people in their daily activities, such as water sprinkling[3], smart thermostats[4] and floor vacuuming and mopping[5]. Moreover, as the domestic environment becomes increasingly instrumented with sensors through the Internet of Things (IoT), it can be expected that these systems will become even more common. Furthermore, smart systems might even become essential to manage the wealth of data sensors generate, such as data mining for hierarchical inhabitant model for controlling a smart environment based on sensor observation (Youngblood and Cook (2007)), relieving their users of significant cognitive and physical workload involved in performing their daily activities. Some examples are: ambience-control[6], iSmart[7], and vocera[8] that control the temperature of a room, ambient lighting, appliances' activation, security, and hot water through different sensors (i.e. temperature sensors, motion sensors, light intensity sensors, and smoke sensors) that control actuators (i.e. lights, plugs,

[1] https://education.smarttech.com/en.
[2] https://smarttech.com/en/support/browse+support/product+index/hardware+products/smart+table/442i.
[3] http://www.rainbird.com/homeowner/products/systems/32ETI.htm.
[4] https://nest.com/.
[5] http://www.irobot.co.uk/home-robots/.
[6] http://ascentech.co/product/ambience-control/.
[7] http://www.ismarthomecontrol.com/.
[8] https://www.vocera.com/.

security cameras, and heaters). In summary, these smart systems have the capacity to automate or reduce people's everyday activities decreasing their workload.

While the potential benefits of smart systems are clear, there are open questions around users' perception of, understanding, and interaction with such systems (Norman (2013)). Some studies have shown that some people have problems understanding how smart systems perform their task or even worse believe that the systems do not perform their task (Kato et al. (2014)). Kato et al. ran a field study with a to-do list interface for sharing tasks between human and multiple agents including software and robots. In this study, they found that one participant was concerned that the robot did not perform its task (i.e. vacuuming the house), even though the robot worked perfectly. This finding suggests that people can perceive that robots do not function properly, regardless of the fact that the robots actually worked. Other studies have shown that people do not understand how systems perform their task or the rules the systems follow to fulfil them (Alan et al. (2016); Forlizzi and DiSalvo (2006); Tullio et al. (2007); Yang and Newman (2012)). Alan et al. (2016) ran two field studies where two different automating energy tariff-switching were developed and evaluated. Both systems offered flexible autonomy[9]. However, people understood the tariff agent system to be smarter than it was (the system was not learning over time, but some participants reportedly perceived it to). Forlizzi and DiSalvo (2006) ran an ethnographic research on the use of Roomba (vacuum cleaning robot) in the domestic environment. From the interviews they conducted with their participants, they found that participants perceived that the robots were not smart enough because they did not quickly memorise their environment. However, Roomba robots do not memorise rooms as people expect, as such, people find it difficult to understand how Roomba robots navigate through rooms. Moreover, Tullio et al. (2007) ran a six-week field study to analyse whether intelligibility can help office workers improve their understanding of how a system predicts their managers' interruptibility. They found that people were able to understand the system prediction better, even if the overarching structure of their mental model stayed stable during the study. Additionally, Yang and Newman (2012) focused on analysing how people understand smart systems (i.e. smart thermostat Nest). They ran a diary study to analyse how people perceive and adapt to the system through the web and their physical interface. In the interviews that Yang and Newman conducted, they found that sometimes there are gaps between user's expectations of the Nest and actual system design. From these findings, we observe that the lack of understanding of how the systems perform their task lead people to misunderstand the system. Hence, people perceive that the systems did not perform well as they expected. As such, as designers, we need to focus on finding an approach that influences people's perception of smart systems.

As Yang and Newman (2012) claim, designers need to ensure that users can understand how their systems work and improve how users perceive them. Hence, they propose that

[9]Systems will sometimes be required to work entirely autonomously, but will often be controlled by users.

systems need to provide information that makes the system intelligible by describing current systems' task. Eslami et al. (2015) showed an implementation of intelligibility as a feedback. In more detail, Eslami et al. ran a user study with Facebook users to observe their perceptions of the Facebook News Feed curation algorithm. For the study, the majority of the users were not aware of the existence of an algorithm to manage the news they received. Their findings suggest that the people's satisfaction levels on the news they received did not increase after they became aware of the algorithm's presence and when they understood how the algorithm worked. However, they found that users interacted more on Facebook and they started to feel more in control of the site. Moreover, Herlocker et al. (2000) ran a study to understand *how* and *why* designers should implement explanation interfaces for automated collaborative filtering systems. They found that it is not straightforward to design meaningful explanations that can explain how the systems fulfil their task because the systems' complex mathematical models. Moreover, they found that the explanation they designed was too complex for non-expert users, while in contrast, expert users understood the explanations. In addition, Lim and Dey (2011b) ran a study with two context-aware systems (HearMe and LocateMe). Their results suggest that the impact of explanations on users impressions depends on the systems' certainty and behaviour appropriateness. Hence, explanations are helpful for systems with high certainty, but it is also harmful if the systems' certainty is low. As such, they propose that designers should be cautious that systems are sufficiently certain before they design and implement explanations on context-aware systems. Additional to these findings, Ehrlich et al. (2011) showed that the implementation of explanations require time and high cognitive workload for users to understand the explanations given. Furthermore, these explanations may even lead to lower decision quality for some users. As such, designers need to find a way to design explanations that can be understandable for any user. In contrast, another approach to improve people's understanding of how smart systems work is information visualisation[10]. As Verbert et al. (2013) claim, researchers have been focusing on the development of algorithms to improve the accuracy of systems' rather than support exploration and control by end users. Hence, they ran a study with a recommender system (TalkExplorer) to analyse how information visualisation can improve people's understanding of how the system generates recommendations. Verbert et al. (2013) implemented TalkExplorer to analyse how people perceived relevance and meaning in the recommendation process. They found that people value the visualisation as a way to gain insight into systems' recommendation and they performed better in finding relevant information that helped them to attend a conference. These findings have demonstrated that the implementation of explanations is a useful approach for some systems.

There is an inherent tension between making the system's operation visible to its users and hiding its complexity (e.g. regarding pattern recognition, artificial intelligence,

[10]Information visualisation is the study of interactive visual representations of numerical and non-numerical data to reinforce human cognition of smart systems behaviour and performance.

or machine learning models). Moreover, the design of adequate explanations can be difficult and requires time, previous knowledge, and effort on behalf of the user to process and understand the information embedded in the explanations (Ehrlich et al. (2011); Herlocker et al. (2000); Lim and Dey (2011b)). Furthermore, there is a potential cost at design time and effort from users. As such, Vermeulen (2010a) proposes a different approach to explain systems' decision. In their study, animations were used to show the process that a system follows when it makes a decision, given how a user interacts with its inputs (e.g., switch) or sensors (e.g., motion detector). Findings from their study suggest that participants understood decisions and actions taken by the system due to the explanations they received. This approach demonstrates that an animation, as feedback, can help people understand a system's decisions. However, participants also found it difficult to track the animation at times, thereby confusing them.

Even though there are studies that have been analysing how to improve people's understanding of how the systems perform their task, there is a research question that needs to be addressed around how to implement a feedback that can modify the perception of users without complicating the systems or increasing user's mental workload. Hoffman and Vanunu (2013) focused on analysing how to change people's perception in an external event that is related to the system's task. However, they used a pet-like robot (*Travis*) that used motion feedback and physical characteristics to enhance how people perceive an external event. In more detail, how people perceive the music that was played by Travis. These findings are relevant for our research because they demonstrated the possibility to implement *visual feedback* to change people's perception without delivering any explanations. However, this study opens a new research question: Can *visual cues* influence how people perceive smart and autonomous systems' performance? In particular, *visual feedback* that is intrinsic to the system (i.e. part of system functionality rather than changing systems' physical design) or requires small implementations or modifications that can trigger the *visual cues* that we want to implement.

In summary, previous studies have demonstrated that the lack of knowledge of how smart systems perform their task affects how people perceive them. Hence, researchers have proposed to make systems' decisions clear and how to perform their task through meaningful explanations. However, providing explanations require time, high mental workload or even worse people find it difficult to understand the explanations. Even if people do not understand how smart systems work or perform their task, designers can change how people perceive smart systems. Hoffman and Vanunu (2013) have shown that it is possible to change how people perceive the systems without delivering explanations and keeping the interaction with the user as simple as possible.

Against this background, in this thesis, we focus on finding an approach that can improve how people perceive smart systems' performance with the implementation of *visual feedback* that is intrinsic to the systems' task. In more detail, first, we focus on analysing how *visual feedback* as a *physical motion cues* affect how people perceive robots' system

performance (i.e. domestic cleaning robots' task). Additionally, we are looking to analyse different approaches that can show *visual feedback* remotely (e.g. streaming videos as a *visual feedback*). As Woods et al. (2006) showed in their study, people were in agreement between live and video-based human-robot interaction trials. In more detail, these results suggest that people can observe equally a robot in a video as a one physically. However, we need to analyse if a video implementation can have the same effect on people's perception as seeing the robot working physically. Second, we want to find a *visual feedback* that shows motion in screen-based systems (i.e. optical character recognition and text processing systems) that can have the same effect as a *visual feedback* that a robot system can show. However, we need to consider that the implementation of motion as a *visual feedback* is not intrinsic to screen-based systems as it is on robot systems, as such, we need to find an approach that shows motion as a feedback without complicating the design of such systems. Moreover, we need to understand what characteristics the *visual feedback* need to have to change how people perceive the performance of the systems. From previous studies (Norman (2013); Park and Lee (2010); Tremoulet and Feldman (2006)), we learned that the characteristics to be analysed in the design could be: detail of movement displayed, a motion that is human-like, or movement that represents people's mental model of how they consider a system perform its task. As Healey and Enns (1999) claim, people detect rapidly and accurately detect features in objects that display movement on a screen independent of the total of amount of elements displayed. However, we need to analyse whether the details of movement displayed affect people's perception. Additionally, Rousseau and Hayes-Roth (1997) and Mateas (1999) suggest that robots should exhibit naturalistic behaviour and in some cases appropriate emotions to affect how people perceive them. Moreover, they found that people require little or no knowledge or effort to understand and interact with the robots. Thus, we consider that the design of the *visual feedback* requires a human-like motion to influence people's perception. Finally, as Norman (2013) suggests, mental models, play a fundamental role in how people interact with the objects and systems. When people have an erroneous mental model of the system, they can find difficulties in using such systems. In contrast, when the mental model is valuable in providing understanding, people can predict how systems will behave. Hence, we consider that people's mental model needs to be discussed in the design to ensure that the design of the *visual feedback* can have an influence on people's perception. All these implementations and designs are analysed in this research to inform designers how they can affect people's perception of the performance of robot and screen-based systems.

1.1 Research Questions

In summary, this thesis attempts to answer the following research question: can *visual cues* influence how people perceive smart and autonomous systems' performance?

In particular, the question is assessed through showing *physical motion cues* for autonomous robots and *animation cues* for screen-based systems. First, we decomposed the question how *physical motion cues* affect how people perceive autonomous robot systems' performance in small ones. These questions are:

- Can *physical motion cues*, which are intrinsic to the systems, influence people's perception of the performance of autonomous robots?
- Do *physical motion cues* are more effective than *video-based cues* at influencing how people evaluate an autonomous robot that shows such cues?
- Can *video-based cues* influence people's perception of autonomous robot's performance?

These research questions are extensions of how people perceive robot systems after they receive a *visual cue* (Hoffman and Vanunu (2013)). Additionally, we focused on understanding how *animation cues* affect how people perceive screen-based systems' performance. Hence, we also focused on answering the following question: what characteristic (e.g. detail of animation, a motion that is human-like, or movement that represents people's mental model of how they consider a system perform its task) these cues need to have to make a positive influence rather than a negative one? These questions are:

- Can *animation cues* influence people's perception to score higher smart system's performance?
- Do *animation cues* have an effect on people's perception of a smart system's performance if the animation of the system processing the task is similar to how a human would process the task?
- Do *animation cues* have an effect on people's perception of a smart system's performance only if the animation is consistent with people's mental model of how the system works?
- Can higher detail of animations better influence people's perception of smart systems than a lower detail of animations?
- What level of imbalance in the performance level of the system being compared would "break the illusion" created by the *animation cues*?
- Does varying the speed of the animations would "break the illusion" created by the *animation cues*?

These research questions will help us to propose an effective approach that can help designers in designing effective *visual cues* (i.e., *physical motion cues* and *animation*

cues) that will affect how people perceive the performance of robot and screen-based systems respectively. In the following section, we present our research contributions that show two studies that talk about the effect we found with the implementation of *physical motion cues* in robot systems.

1.2 Research Contributions

This thesis resulted in the following contributions. In more detail, these are the following research contributions that we present in UbiComp'16:

- We present two lab studies designed to investigate whether showing *physical motion cues*, which is showing the process of a system through movement (that is intrinsic to the system's task), of a vacuum cleaning robot as it completes its task, affects how users perceive its performance. Our results suggest that physical presence does yield higher performance ratings.
- We present findings from one lab study and seven follow-up studies on the crowdsourcing platform designed to investigate the potential of *animation cues* to influence users' perception of two smart systems: a handwriting recognition and a part-of-speech tagging system. Results from the first three studies indicate that indeed *animation cues* can influence participant's perception of both systems' performance. The subsequent three studies, designed to try and identify an explanation for this effect, suggest that it is related to the participants' mental model of the smart system. The last two studies were designed to characterise the effect more in detail, and they revealed that different amounts of animation do not seem to create substantial differences and that the effect persists even when the system's performance decreases, but only when the difference in performance level between the systems being compared is small.

The research presented in Chapter 3 was also published in the following full papers at international conferences:

> Pedro Garcia Garcia, Enrico Costanza, Sarvapali D. Ramchurn, and Jhim Kiel M. Verame. 2016. The potential of *physical motion cues*: changing people's perception of robots' performance. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16).

Finally, a journal paper about the work presented in Chapter 4, 5, and 6 is currently under review:

Pedro Garcia Garcia, Enrico Costanza, Jhim Kiel M. Verame, Diana Nowacka, and Sarvapali D. Ramchurn. 2017. Seeing (Movement) is Believing: the Effect of Motion on Perception of Automatic Systems Performance. In, Human-Computer Interaction (under review)

1.3 Thesis Structure

The remaining chapters of this thesis are structured as follows:

Chapter 2 – Related Work – The chapter starts by reviewing transparency for the intelligibility of software and robotic systems, which focuses on delivering meaningful explanations describing systems' behaviour. An overview of research related to the perception of motion in screen-based systems, robots, and interactive artefacts, followed by a discussion of the perception of robots motion through video and animation. The chapter concludes with references to cognitive bias to analyse what makes people take a decision or persuade them.

Chapter 3 – How Physical Motion Cues Change People's Perception of Vacuum Cleaning Robots Systems' Performance– The chapter presents three laboratory studies that were designed to test the effect of *physical motion cues* in robots systems. We used a vacuum cleaning robot (Roomba) to implement the *physical motion cues* that we presented to our participants. To this end, we compared three conditions: (i) *motion* condition, participants saw the robot moving as it docks into its charging station in person, (ii) *video* condition participants saw the robot docking through a video-based notification, and (iii) *no-motion* participants saw the static robot that has already docked to the charging station.

Chapter 4 – How Animation Cues Change People's Perception of HWR and POS Screen-based Systems' Performance – This chapter reports one laboratory study and two crowdsourcing studies to analyse the effect of *animation cues* in smart systems. We implemented a Handwriting Recognition system (HWR) and Part-of-Speech tagging system (POS) to test the effect of *animation cues* on people's perception of screen-based system's performance.

Chapter 5 – What Makes Animation Cues Affect People's Perception? – This chapter summarises three crowdsourcing studies that were designed to address the question why *animation cues* influence people's perception of systems' performance. We tested whether the animation designed could be considered as for how a human can perform the task, or because the animation matches people's mental model are the reason behind the effect of *animation cues*.

Chapter 6 – The Effect of Varying Other Dimensions of the Animation Cues – This chapter reports three crowdsourcing studies that focus on analysing the varying of other

dimensions of animation (e.g. amount of details, the number of errors, and speed). In more detail, the first study focuses on testing how much animation needs to be shown in all the elements related to the system's task, for it to have an impact on how people perceive the performance of screen-based systems. Additionally, the second study focuses on observing for how long the positive effect of *animation cues* on people's perception persists even the system's performance starts to degrade. Finally, the third study analysed whether the varying of *animation cues'* speed can be a factor that can change how people perceive system's performance.

Chapter 7 – General Discussion and Conclusions – The last chapter summarises and discusses the main achievements of this thesis. Furthermore, the implications design for future implementations of *physical motion cues* and *animation cues* as *visual feedback*. Finally, we also summarise the thesis in this chapter.

Chapter 2

Related Work

Our research aims to analyse how *visual cues* through physical motion in autonomous robots and animation in screen-based systems can change people's perception of smart systems' performance. To this end, in what follows, we survey prior research that has studied cognitive biases and how the different framings of information impact people's perception. Then, we discuss transparency for the intelligibility of software and robotic systems; The role of motion in users' perception of robots, interactive artefacts and screen-based systems; and perception of robots motion through video and animation.

2.1 Transparency and Intelligibility of Software Systems

Prior research has examined the effect of increased intelligibility on people's understanding of smart systems. In particular, previous studies have suggested that smart systems should generate and provide meaningful explanations for systems' actions, behaviour or outcomes (Lim et al. (2009); Lyons (2013); Tullio et al. (2007)). For example, Lim et al. [Ibid.] ran two experiments to analyse the effect of meaningful explanations describing *why* and *why not* a context-aware application behaved in a certain way. Their findings suggest that users have a better understanding of a system's behaviour and a higher feeling of trust in it when it provides explanations. Moreover, Tullio et al. (2007) ran a six-week field study to analyse whether intelligibility can help office workers improve their understanding of how a system predicts their managers' interruptibility. They found that people were able to understand the system prediction better, even if the overarching structure of their mental model stayed stable during the study. However, explanations can also cause information overload, possibly confusing and overwhelming users (Lim and Dey (2011a); Yang and Newman (2013)). Similarly, another study investigated *Laksa*, a context-aware software which used eight question type explanations (e.g. *Why, Why Not, What If*) to explain its decision to the users (Lim and

Dey (2011a)). To evaluate the software, participants used the software in three situational dimensions (exploration, fault finding, and social awareness) that allowed the researchers to observe whether participants do or do not understand software decisions. They noted that quickly consumable explanations of a system's output are crucial and additional, richer explanations should be easily accessible. Lim and Dey [Ibid.] observed that prior knowledge plays an important role in both understanding of such systems and also interpreting the explanations given. The lack of previous knowledge can lead people to misunderstand or misuse a system. Complementing this prior work, our aim is to understand whether it is possible to change people's perception of smart systems without increasing their cognitive workload by, for example, providing additional cues (e.g. through animation) that can expose to users that a smart system is doing work.

Another way of improving system intelligibility is through information visualisation, which is the use of visual representations of data structures and algorithms to help people analyse data (Card et al. (1999); Ware (2012)). The concept of information visualisation is considered a method to make a system understandable without providing explanations of its process. For example, O'Donovan et al. (2008) ran a study where participants interacted with *PeerChooser*, an interactive visualisation system for collaborative filtering. The system generated a peer-graph which is centred on the current user. The graph showed a visual representation of their peer group or neighbourhood allowing participants to manipulate connections with their neighbours. This interaction allowed participants to visualise recommendations from the system based on their preferences. Their findings suggest that a visual-interactive approach can improve the accuracy of the recommendations provided by the system and also enhance user experience (O'Donovan et al. [Ibid.]). In our case, instead of using interactive visualisations, we explore visualisations of a system's process through motion (animations) that represents its execution of a task.

An example of a study that uses motion as a *visual feedback* to explain a system's decision is presented by Vermeulen (2010b). In their study, animations were used to show the process that a system follows when it makes a decision, given how a user interacts with its inputs (e.g., switch) or sensors (e.g. motion detector). Findings from their study suggest that participants understood the decisions and actions taken by the system because of the explanations they received. This approach demonstrates that animation, as a feedback, can help people understand decisions made by a system. However, participants also found it difficult to track the animation at times, thereby confusing them. We build on this idea and want to explore further how people's perception changes depending on the animation.

2.2 Transparency and Intelligibility of Robots

Robots are part of people's everyday life in homes and public areas (e.g. in hotels, trade shows, workplaces, museums) (Schulz et al. (2000)). Therefore, robots should be transparent about their decisions and actions so that people would feel that they understand their behaviour (Kim and Hinds (2006)).

In the study by Kim and Hinds [Ibid.], they examined the impact that different levels of transparency had on people's judgment of autonomous robots. For the study, they made use of *Pyxis HelpMate*, a robot that can deliver medication in the hospitals. In particular, the robot provides audible feedback about its status when it malfunctioned. However, they did not focus on understanding how people perceive the performance of the robot with the different levels of transparency that the robots provided through the audio feedback. Their interest was to understand whether users blame themselves, other users or the system itself in a situation where the robot malfunctions and provide different levels of transparency through an audible explanation. They found that transparency allowed participants to attribute responsibility to others rather than the robot. Boyce et al. (2015) ran an experiment with a simulated autonomous robotic squad member. They implemented an external interface (screen display) to display the robots and the different level of transparency of the robots. Their results showed that increasing transparency could help users understand a robot's environmental conditions and status. In contrast to both of these studies, we do not enhance the existing structure of robots. Instead, we utilise their existing (unaltered) setup as a way to keep the design of the robots as simple as possible.

2.3 The role of motion in users' perception of screen-based systems

Research has looked at how people perceive motion in screen-based systems. *Animacy*, as Tremoulet and Feldman (2000) state, is when people perceive an object as being alive, through the pattern of its movements. They mention that the movement of an object does not need to be dramatic to show animacy (Fritz Heider (1944); Reeves and Nass (1996)). As a consequence, people attribute motivations or intention in objects' movements from the patterns that these objects follow. This means that people can infer objects' intentions through their movements (Gao and Scholl (2011); Michotte (1963); Pantelis and Feldman (2012); Schlottmann and Surian (1999)). This has also been observed during people's interaction with physically actuated interfaces such as helium balloons (Nowacka et al. (2015)). Therefore, through designing the movement, it is possible to affect how people perceive objects. Michotte (1963) showed in their study that if two objects are in the same frame and suddenly change their direction, people

can infer that both objects have a causal interaction. Pantelis and Feldman (2012) ran a study with multiple objects moving around on a screen. They found that, after watching multiple objects moving on a screen, people make interpretations of the intention and behaviour of the objects. Moreover, in their experiment, people were able to distinguish if an object behaved friendly or hostile when it was moving around other objects. This body of work makes us believe that - by showing people an animation - they can be convinced that a system is working on a task. As such, we presume that people perceive a system that somehow communicates that it is doing work perform better than a system that hides how it works.

However, it has also been shown that some features of animations can confuse people and negatively impact people's perception. These features include but are not limited to: interaction between multiple objects (Gao et al. (2010)), trajectories that are too complicated (Dittrich and Lea (1994); Tremoulet and Feldman (2000)), unnatural movements (Popović et al. (2003)), or static backgrounds that are too complex (Gelman et al. (1995)). Hence, it is important to ensure such issues are avoided when providing feedback about a system's execution of tasks.

Prior research has also analysed affective qualities of an interface depending on how the information and motion are presented on a screen (Detenber and Reeves (1996); Park and Lee (2010)). Park and Lee [Ibid.] ran a study to understand how motion (i.e. transition effects between objects) influences the affective quality of an interface to improve user experience. They presented an image viewing interface that allows users to browse through a set of photos as they shift horizontally from one to another. Their results show that motion influenced how people rated affective qualities of the interface (e.g. youthfulness, calmness, and uniqueness). Also related to the effect of animation on user emotion, Bakhshi et al. (2016) reported that social network users have a tendency to share content more frequently if it involves animations, compared to content that is purely static. In contrast to this prior work, our interest lies in observing if motion has an effect on how people perceive systems' performance rather than on people's emotions.

2.4 The role of motion in users' perception of Robots and Interactive Artefacts

Prior studies in HCI, HRI, and UbiComp have examined whether people can infer intentionality, emotions or are motivated to interact with robots or artefacts through visualisation of motion (Bretan et al. (2015); Dragan et al. (2015); Jung et al. (2013a,b); Mortensen et al. (2012); Nowacka et al. (2015)). Mortensen et al. [Ibid.] run a lab study with a motorised TV. The aim of the study was to analyse if motion can communicate agency related attributes such as social status or likeability. The experiment

was between-participants; five participants interacted alone with the TV and ten participants interacted with the TV in pairs. Their results show that some participant felt that the motorised TV "liked" them. However, some participants did not like that the movement of the TV was autonomous. On the other hand, they observed that some participants felt that the TV was rude to them when it follows only one participant in the pair condition. In the study by Nowacka et al. (2015), they tested an actuated helium balloon *Diri*, which interacted as a social agent. For the study they developed two systems: Diri #1 was designed to move randomly in the environment, and Diri #2 was autonomously moving around its environment avoiding obstacles. For the study, they ran a workshop where the participants interacted with the robots. Their results suggested people try to make sense about the behaviour and movement of autonomous systems (e.g. avoiding obstacles). Moreover, they found that people associate mechanicals systems with animal forms, especially to pets. Additional to these findings, Jung et al. (2013a) ran an experiment with three prototypes of moving products (*water-dock, assignment box,* and *recycle-bin*). The aim of these interactive artefacts was to conduct user experience field studies to test the effect of motion in the visceral, behavioural, and reflective perspectives. They found that people react to the movements of the objects, making them interact more with the products. Moreover, people found intentions on products' movement, and they made sense of anthropomorphic meaning from the movements. These studies aimed to understand which emotions were triggered when people see artefacts in motion. Furthermore, their results suggest that motion makes people link artefacts with living beings. Instead, in our study, we use the motion of a robot as a *visual cue* to change how people perceive the robot's performance.

Closer to our work, Hoffman and Vanunu (2013) conducted a study where a pet-like robot, *Travis*, was used as a speaker dock and music listening companion. Participants observed, listened and evaluated songs played by Travis. For some participants, Travis moved on-beat with the songs played. In contrast, other participants interacted with a moving Travis, that was off-beat with the songs. The rest of the participants were introduced to a static Travis. Their results showed that participants rated songs significantly higher when the robot is moving on-beat with the songs than the other two conditions. Indeed, they pointed out the role of "personal robots as contributors to, and possibly amplifiers of, people's evaluation of external events".

These findings focus on the evaluation of events that are external to the system e.g. asking people whether they enjoy what they hear, instead of asking them the quality of the sound produced by the system. In contrast, we focus on how people evaluate automatic systems' performance. Moreover, they centred their research on pet-like robots. Instead, we are particularly interested in everyday systems (e.g. systems that are used in everyday situations such as cleaning robots). However, it would be neither practical nor feasible to reinvent everyday systems to be anthropomorphic. As such, we focus on maintaining the simplicity of such systems.

2.5 Perception of robot motion through video and animation

Previous studies showed how people perceive robots through their physical movement that is related to robots' task (Hoffman and Vanunu (2013)). However, there are other alternatives to interact with robots, such as videos and animations. Such modalities allow people to visualise robots remotely without physical interaction with the system. Takayama et al. (2011) examined people's perception of virtual animated robots through a lab study. For the study, the robot covered a variety of activities, such as opening a room door, delivering a drink, requesting help from a person to plug into an outlet, and ushering a person into a room. Their results suggest that people are positively influenced by animations showing the outcome of robot's task, and more specifically that they read robot's behaviour with more certainty. However, while the focus of their study is on training, we are interested in real-time interaction with robots. Additionally, while their work is based on virtual animated robots, ours use physical ones. Wainer et al. (2007) ran a study with participants that interacted in a collaborative task with an embodiment robot[1] vs. Non-Embodiment robot (e.g. simulated and video). In more detail, participants resolve a Towers of Hanoi puzzle following the instructions of the robots. Their results suggest that people perceive an embodiment robot to be more helpful and enjoyable in comparison with a non-embodiment robot. However, they did not analyse whether people perceive that one type of robot works better than the other, which is instead our key contribution.

2.6 Cognitive Biases

Studies in psychology and behavioural economics have shown that people's perception of how well a system or process works can be influenced by different cognitive biases. For example, Tversky and Kahneman (1985) showed that people could be influenced by the way outcomes are described to them. In a survey, participants were presented with a problem and two possible solutions. These two solutions had the same outcome, however, one emphasised its positive aspects, while the other emphasised the negative aspects. Results suggest that people had a tendency to choose the solution that emphasised the positive aspects. As another example, Ariely (2008) ran a study to analyse if the price on medicine has a placebo effect on people's perception of how they feel after they took medication. One group received the medicine with the actual price and a second group received the medicine with a 10 cents discount (off an original price of $2.50). The results showed that while almost all participants in the first group experienced pain relief from the pill, *only half* of the participants who were given the "discounted medicine" experienced pain relief. While such studies motivate our work, none of them

[1] Embodiment requires a coherent physical realisation to persist over time.

has looked at the whether different framing, through *visual cues*, can also influence people's perception of smart systems.

Our interest in using motion as a *visual cue* to change people's perception of system performance is motivated by a large body of work from cognitive psychology, which investigated how motion and other sensory cues influence our perception of the world. In psychology "perception" is defined as the process that people follow to identify, interpret, and understand their environment, with the support of sensory (i.e. physical) and cognitive cues (referred to as *high-level of knowledge*) that the nervous system processes (Schacter et al. (1978)). Studies have shown that humans can extract high-level information from very basic motion cues (Johansson (1973)). However, in some cases, physical cues are insufficient for the brain to interpret the environment. Hence, the brain uses existing knowledge as a way to make sense of sensory signals (e.g. sight) (Richard (1998)).

Our perception of the world is sometimes influenced by more than one sensory channel. For example, McGurk and MacDonald (1976) demonstrated that both sound and vision influence speech perception. Vines et al. (2006) reported a study where participants rated how much they liked audiovisual clips of clarinettists, to investigate how different *visual cues* affect people's evaluation of the musicians' performance. They found that participants gave a lower score to a clarinettist who did not move compared to a clarinettist with more expressive body motion. This result suggests that an appropriate *visual cue* can improve people's evaluation of something that is not visual. Building on such a corpus, we set out to explore whether motion can be leveraged to influence people's perception of smart systems.

2.7 Summary

Having analysed the various existing frameworks in detail in the previous sections, we now proceed to present a summary that highlights what has been archived in the field of transparency and intelligibility of smart systems, how users perceive the motion of smart systems, and cognitive bias. In this way, we aim to identify and propose a modality that can affect users' perception of smart systems.

In summary, previous studies have shown that people's perception of how the systems perform their task can be affected because of the lack of knowledge that they have about systems' behaviour. Thus, researchers have proposed that automatic system should generate and provide meaningful explanations for systems' actions and behaviour (Lim et al. (2009); Lyons (2013); Tullio et al. (2007)). However, researchers have noted that prior knowledge plays a major role in supporting people's understanding of how the systems work and also interpreting the explanations that the systems generate (Lim and Dey (2011a)).

Against this background, Vermeulen (2010b) showed that the implementation of motion as a *visual feedback* to explain systems' decision is an approach that can support users' understanding of how systems perform their task. In particular, Vermeulen proposes an animation that showed the process that a system follows when it makes a decision. However, users can find difficult to follow the animation at times, thereby confusing them. We are building on this idea and want to explore further how *visual cues* (*animation cues* and *physical motion cues*) as a feedback affect people's perception rather than use *visual feedback* to explain system's decision. Hence, we focused on analysing the role of motion in users' perception of smart systems.

Building on this prior work, Hoffman and Vanunu (2013) have shown that it is possible to change how people perceive the systems without delivering explanations and keeping the interaction with the user as simple as possible. In more detail, they found that people rated significantly higher a robot system's outcome when this one shows movement while is performing its task (i.e. movement is not related to main robot task). However, our focus is on how people perceive the performance of smart systems (e.g. robot and screen-based systems). Additionally, we are particularly interested in everyday systems (e.g. systems that are used in everyday situations such as cleaning robots). Hence, we focus on keeping the simplicity of the systems, such as, using robots' movement or implementing small animations that represent systems' task for screen-based systems.

In addition, we consider that the implementation of *visual cues* as feedback can influence people's decision-making. As such, we found different studies in psychology and behavioural economics that have shown how people's decision can be influenced by different cognitive biases (Ariely (2008); Tversky and Kahneman (1985)). However, none of them has analysed whether *visual cues*, can also affect people's perception of smart systems.

Chapter 3

How Physical Motion Cues Change People's Perception of Vacuum Cleaning Robots Systems' Performance

We designed three studies to observe the effect of *physical motion cues* on people's perception of vacuum cleaning robots systems' performance. To this end, we used a Roomba robot to test how people react once they observe briefly a robot system docking (i.e. robot moving to its base station). In the following chapter, we explain the design of the three studies we conducted to address our first three research questions. In more detail, Study 1 (N=16) was designed to address the question " Can *physical motion cues*, which are intrinsic to the systems, influence people's perception of the performance of autonomous robots?". In addition, Study 2 (N=16) was designed to address the research question "Do *physical motion cues* are more effective than *video-based cues* at influencing how people evaluate an autonomous robot that shows such cues?". Furthermore, we designed Study 3 (N=16) to address the third research question "Can *video-based cues* influence people's perception of autonomous robot's performance?". Finally, we present and discuss our findings to understand their implication for design.

3.1 Study 1 – Physical motion cues vs. no-motion

We selected the Roomba robot because it is an off-the-shelf product designed for everyday domestic use. The aim of using this robot was to help us to analyse whether *physical motion cues*, which are intrinsic to the system, can influence people's perception of the performance of autonomous robots.

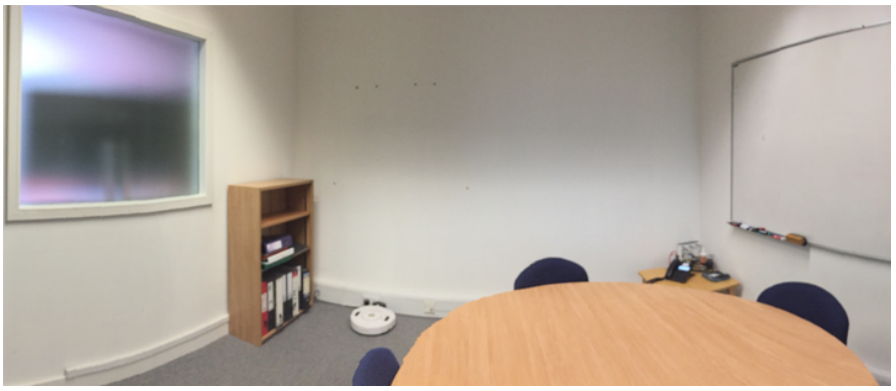

Figure 3.1: Picture of one room where the Roombas worked in their task.

3.1.1 Method

3.1.1.1 Study Design

In our first study, a fully counterbalanced, a within-participants design was used to compare the effect of *physical motion cues* and its absence on the same set of participants. Participants evaluated and compared the performance of two Roomba robots in vacuuming the carpets in two rooms where each robot was located. Two conditions were defined in our study: *no-motion* and *motion* condition. In more detail, for the *motion* condition, the robot showed movement. This movement is the *physical motion cue* at the centre of our study. In practical terms, the *motion* condition was implemented through a Wizard of Oz approach whereby an experimenter activated the robot seconds before participants arrived.

Additionally, we designed a *consensus-oriented reward system* to try and ensure that participants would provide a significant and thoughtful evaluation when they choose which system they considered to have the best performance. We told the participants that the majority who selected the system to has the best performance they will be rewarded at the end of the study with a £10 voucher. Moreover, external validity was a key factor. Therefore, we were particularly careful in keeping a number of variables that could affect participants' perception of the performances of the Roombas constant. These variables were determined through pilot studies:

- Cleanliness of carpets: The Roombas did not actually clean the carpets during the study.
- Robot's environment: The rooms used in the study were similar to maintain the same conditions (see Figure 3.1). Moreover, the robots were switched between the two rooms to maintain a fully counterbalanced study design.
- Roomba task completion time: Both of the Roombas were simulated to vacuum the room in 10 minutes and were working simultaneously.

- Robot's model: The robots that we used for the study was iRobot Roomba 500 for both conditions.
- Evaluation time: Participants were allowed only 15 seconds evaluating the carpets in each room (avoiding that participants spent more time in one room than the other).

3.1.1.2 Participants

A total of 16 participants (12 female, 4 male) took part in the study, and 15 of these were members of the University: PhD and Masters students, none of which had a technical background (e.g., not from Computer Science or Engineering). One participant was a homemaker. The ages of these participants ranged from 24 to 53 years old (M = 32.00, SD = 7.46).

3.1.1.3 Equipment

The study was run in a lab environment, where two iRobot Roomba 500 were installed in two similar rooms. We developed a web page to display the evaluation questionnaire and a puzzle game[1] in a 13" laptop.

3.1.1.4 Procedure

At the beginning of each study, participants were told that the task was to compare two different algorithms implemented on each of the two robots. After this introduction, participants were asked to visit two rooms and were given a questionnaire asking them to evaluate on a 5-point Likert scale the *cleanliness of the carpet* (from "1 - dirty" to "5 - clean without any chance to improve"). They were then asked to move to a different room to wait until the Roombas finished vacuuming the carpets. Participants were explained that they had to wait in a different room because the algorithms were still work-in-progress so we did not want their judgement to be influenced by the robots' trajectories. Figure 3.2 shows the layout of the room. As we mentioned in the subsection study design, two conditions were designed for the study. In more detail, in the *motion* condition (the 'treatment' condition) participants saw the robot moving as it docked in its charging base, having completed its cleaning duties. Half of the participants saw first the robot in the *motion* condition, vice versa for the other half. Moreover, the rooms and the robots were also alternated: half of the participants saw the robot in the *motion* condition in Room A and the other half saw the robot for the same condition in Room B. The first robot participants saw was referred to as simply 'robot A', and the other

[1] http://www.kongregate.com/games/Gibton/blocks.

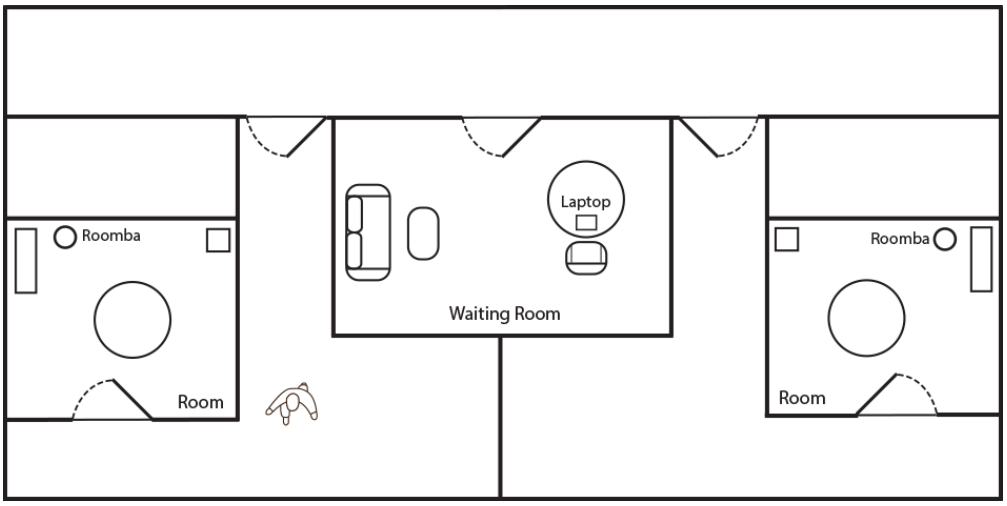

Figure 3.2: This figure shows a layout of the rooms where we conducted the first study.

one 'robot B', regardless of the condition (so that the naming would not influence the results).

While participants were waiting, they were asked to play the puzzle game that we chose for them to keep participants busy while they waited for the robot to finish. When the two robots had completed their tasks, participants received a text-based notification shown on the laptop indicating that they could go and evaluate the performance of the Roombas. After receiving the notification, participants visited both rooms one after the other. As described above, in one room they found that the Roomba has already docked, while in the other room they saw the Roomba docking. After participants had seen the robot docking, we told them that the robot's action of docking was not related to the robot's task of cleaning the carpet. As such, they were allowed to see this part of the robot's process.

After visiting each room, they were asked to continue the questionnaire and evaluate whether there is an *improvement in the cleanliness of the carpet* on a 5-point Likert scale (from "stayed the same, did not have an improvement" to "better than before"). The post-task question was phrased differently from the pre-task questions, so the answers cannot be directly compared. Once they evaluated both rooms, participants were asked to compare the performance of the Roombas. Participants were asked which of the Roombas they thought most people would select the one with the best performance (including the option that both Roombas performed at the same level). This question was the one we designed for the *consensus-oriented reward system*. To check whether participants subjective judgement of the Roombas differed from what they expected the majority of people to choose. After they answered the first question participants were presented with a second question, asking them which robot they consider to be the one with the best performance, regardless of other people's opinion, and the reward. Indeed,

this second question (referred later as a *non-reward-based question*) had no effect on the reward received by the participants.

3.1.2 Results

3.1.2.1 Selection of robot with the best performance.

For the *reward-based* question, 15 out of the 16 participants selected the moving robot (*motion* condition) as the one with the best performance. The remaining participant selected the robot in the *no-motion* condition as the best performing one, while nobody indicated that the robots had the same level of performance. For the *non-reward-based* question, only one participant expressed a different opinion from that of the previous question, saying that both robots had the same performance. In total, 14 participants considered the moving robot as the better performing robot when answering the *non-reward-based* question. These results are illustrated in Figure 3.3.

3.1.2.2 Evaluation of the cleanliness of the carpets.

A Mann-Whitney Test revealed a statistically significant effect ($U = 67.50, p < .05, r = .41$) of the motion on the rating of how clean the rooms were after the operations of the robots. The room was rated on average as cleaner in the *motion* condition ($mdn = 2.5$) than in the *no-motion* condition ($mdn = 1.5$). Figure 3.4 shows participant's evaluation of the cleanliness of carpets. No statistically significant differences were found in the ratings of how clean the rooms were before the operations of the robots between the conditions.

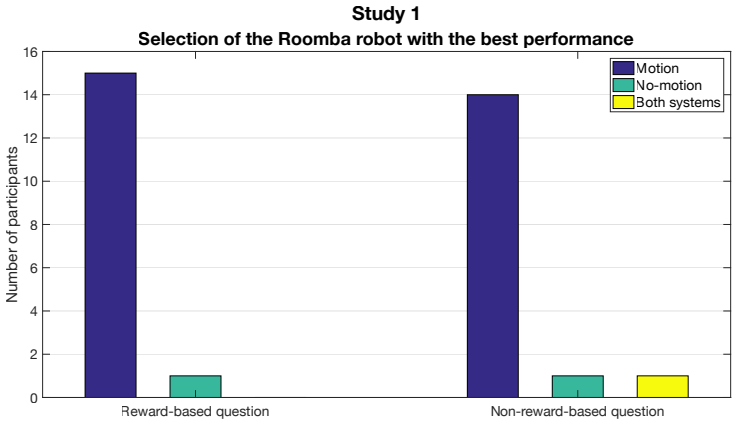

Figure 3.3: Selection of the participants for preferring a Roomba for the *reward-based* question and *non-reward-based* question in Study 1. Number of participants on the y-axis.

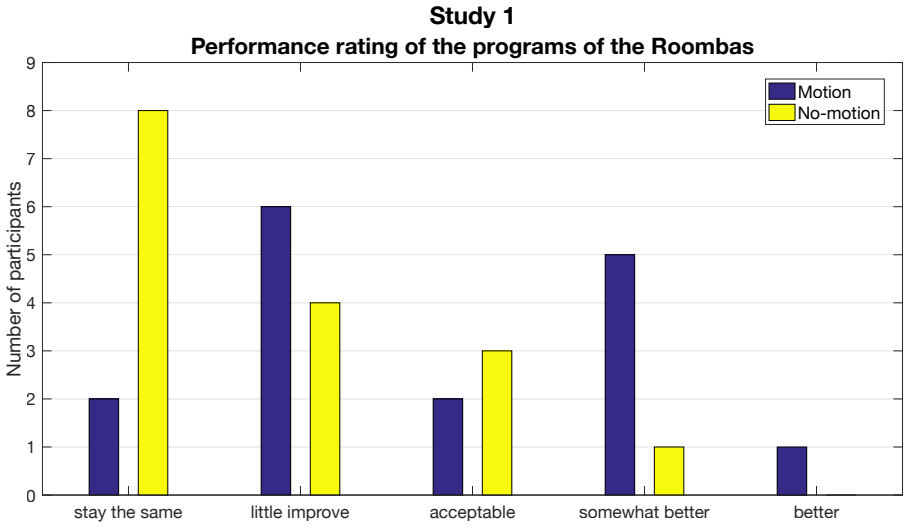

Figure 3.4: Likert-scale of participants' evaluation of the cleanliness of carpets for the *motion* and *no-motion* conditions. Number of participants on the y-axis.

3.1.2.3 Discussion

The results of Study 1 confirm that *physical motion cues* can change people's perception of how they perceive autonomous robots' performance. The data shows clearly that motion cues change how people perceive such systems: all except two participants agreed that the robot in the *motion* condition was the one with the best performance. All except 1 declared that they thought most people would consider the moving robot like the one with the best performance. This finding is further confirmed by how the participants rated the performance of the robots through the Likert scales. In the *motion* condition, the room was rated as cleaner than in the *no-motion* condition, after the robots' operations. As expected, no differences were found on the rating of the rooms before the robots' operations. These results show that motion can be used to change people's perception in the performances of autonomous robots, even in the case of systems that are not anthropomorphic, extending what was previously reported in the literature (Hoffman and Vanunu (2013)).

These results open up a new question, Do *physical motion cues* are more effective than *video-based cues* at influencing how people evaluate an autonomous robot that shows such cues? As such, we designed a second study to analyse whether a video feed has a higher or equal effect as *physical motion cue*.

3.2 Study 2 – Physical motion cues vs. video-based cues

The results from the first study clearly show that seeing a robot moving had an effect on our participants' perception of its performance. However, it should be noted that the

movement was seen *in person*. Do *physical motion cues* are more effective than *video-based cues* at influencing how people evaluate an autonomous robot that shows such cues? Indeed, there might be situations in which users are unable to see the movement of a robot directly. To answer this question, we designed a second study to compare how people perceive the performance of a Roomba when people watch a video of it docking in comparison to watching a Roomba docking in person. As such, we want to analyse whether *physical motion cues* are more effective than *video-based cues* at influencing how people evaluate an autonomous robot that shows such cues.

3.2.1 Method

3.2.1.1 Study Design

The design of Study 2 is the same as Study 1, except that the *video* condition replaced the *no-motion* condition. In the *video* condition, the notification that participants received when the robot completed its task included a video of the robot docking, which the participants saw. *Video* condition is similar to the *motion* condition but mediated over video, rather than physically observed in the same environment. This notification was named as *video-based* notification for further identification.

3.2.1.2 Participants

A total of 16 participants (10 male, 6 female) took part in the study, and all of them were members of the University: undergraduate and postgraduate students, including a wide range of disciplines, from Computer Science to English literature, to Mechanics, to Economics and Psychology. The ages of the participants ranged from 19 to 37 years old (M = 22.00, SD = 4.39).

3.2.1.3 Equipment

The equipment used in this study was the same as Study 1. We used a 13" laptop to display the same web page we designed for the Study. However, we added a *video-based* notification instead of the text-based notification for the *video* condition we used for this study.

3.2.1.4 Procedure

The procedure of this second study was similar to Study 1. However, this study differs that instead of implementing the *no-motion* condition, we implemented a *video* condition to test our second research question. In this new condition, a video of a Roomba docking

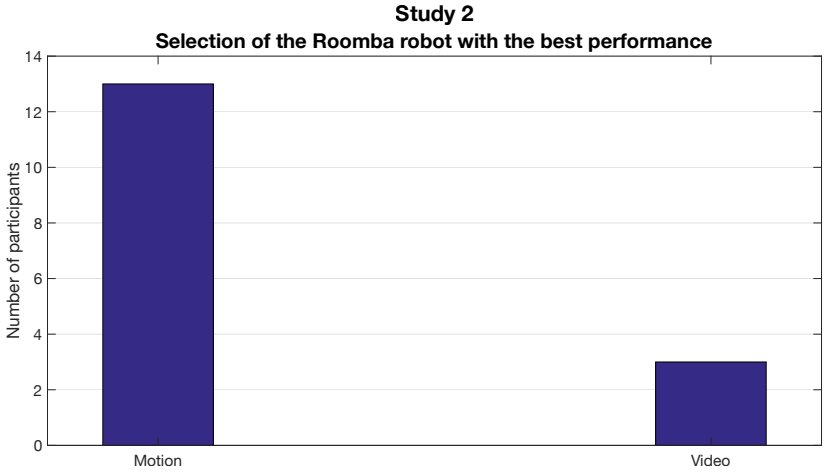

Figure 3.5: Selection of the participants for preferring a Roomba for the *reward-based* question and *non-reward-based* question in Study 2. Number of participants on the y-axis.

was displayed on the laptop computer where the participants played the video game (cfr Study 1), and this served as a notification of this Roomba having completed its operation, rather than the text-based notification. It is worth emphasising that after watching the video participants inspected the room in person. Furthermore, the notifications were shown separately, and each of them was shown to the participants right before they visited the corresponding rooms. This was done to help participants link the notifications to the correct Roombas. For practicality, the video was a pre-recorded clip, but it was presented to participants as a live feed from the room (the two rooms had no external windows, making such mockup realistic). Moreover, the video was recorded in low resolution (VGA) to avoid details that can change people's perception.

We included some new questions in the final questionnaire. In addition to the *reward-based* and *non-reward-based* questions, participants were also asked why they think one Roomba performed better than the other, in order for us to understand the motivation behind their choice. Moreover, they were asked whether they would prefer watching a video of the Roomba working or watching the Roomba physically finishing its task and why.

3.2.2 Results

3.2.2.1 Selection of Roomba robot with the best performance.

In total 13 of the 16 participants considered that the Roomba in the *motion* condition provided a better performance. The remaining three participants indicated that the Roomba in the *video* condition performed the best, while nobody suggested that both

robots had the same performance. All participants gave the same answer to the *reward-based* and the *non-reward-based* questions, i.e. they all believed their answer would be the most popular one. These results are illustrated in Figure 3.5.

3.2.2.2 Reasons for choosing one Roomba over the other.

The responses to the question about why participants selected a particular Roomba as the one performing the best were summarised through thematic analysis (Braun and Clarke (2006)). Each response was associated with one or two themes, with five themes used in total: *details, relative, generic, room features,* and *clean already.* Figure 3.6 illustrates the frequencies of these themes for those who preferred the *motion* condition and those who preferred the *video* condition. The theme *details* was associated with responses which referred to specific issues in the room, such as "crumbs which lie close to chair legs" and "coffee stains." The theme *relative* was used when the responses referred to the comparison of how clean the room was before and after the operation of the robots, such as: "Found the room cleaned by Roomba A much cleaner than it was initially" and "biggest change in cleanness". Comments categorised as *generic* included "cleaned the room better" and "The carpet of room B was cleaner than room A". The theme *room features* was used when participants referred to the influence of room features on the performances of the robots, such as "fewer corners for Roomba to have difficulty with" and "It seemed to clean tighter spaces better". Finally, one participant stated that the room was clean to start with ("Because the Room B is clean already so it is hard to evaluate the Roomba B performance") so this response was categorised as *clean already.*

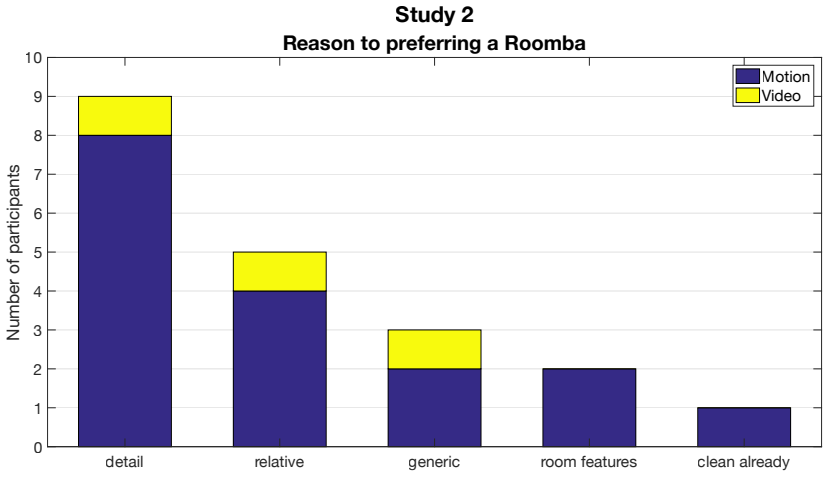

Figure 3.6: Reasons expressed by participants for preferring one Roomba over the other in Study 2.

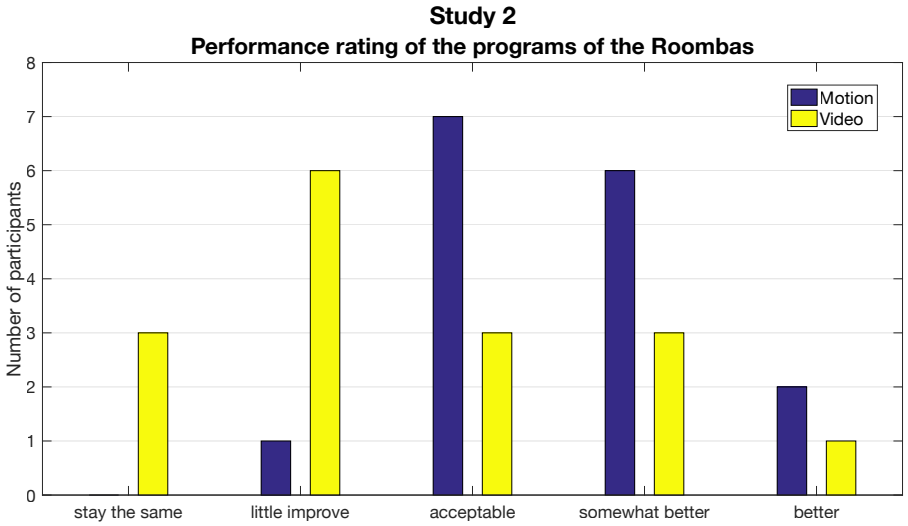

Figure 3.7: Likert-scale of participants' evaluation of the cleanliness of carpets for the *motion* and *video* conditions. Number of participants on the y-axis.

3.2.2.3 Evaluation of the cleanliness of the carpets.

A Mann-Whitney Test revealed a statistically significant effect ($U = 70, p < .05, r = .39$) of motion and video on the rating of how clean the rooms were after the robots' operations. The room was rated on average as cleaner in the *motion* condition ($mdn = 3$) than in the *video* condition ($mdn = 2$). Figure 3.7 shows participant's evaluation of the cleanliness of carpets. No statistically significant differences were found in the ratings of how clean the rooms were before the robots' operations between the conditions.

3.2.2.4 Modality preference.

Ten participants preferred the video overseeing the robot physically move; four participants preferred seeing the robot in person; while the remaining two participants did not have a preference for how they see the robot. Figure 3.8 shows participants' preference.

3.2.2.5 Reason for preferring a modality.

The responses to the question about why participants selected one modality over the other were summarised through thematic analysis. Each response was associated with one theme, with six themes used in total: *better understanding, convenience, emotional, generic, reliable* and *subjective*. Figure 3.9 illustrates the frequencies of these themes. An example in the theme *better understanding* included "I can understand which part of the room has been cleaned". The theme *convenience* included "I do not have to be there until the end", "Can observe the room situation remotely", and "This will save

our time while we are doing some other work during the time Roomba was doing its task [...]". Comments categorised as *emotional* included "fun" and "[...] physical presence has a more personal effect". Comments in the *generic* theme included "You can see the Roomba working physically, and the video is helpful" and "Able to see the functionality of the Roombas". An example comment in theme *reliable* included "On the video you cannot see what is happening". Finally, the theme *subjective* included "I am personally a visual person, so it illustrates it much better...".

3.2.3 Discussion

The results of Study 2 suggest that seeing the robot motion in person influences people's perception on autonomous robots, compared to seeing it through video. All except three participants agreed that the robot in the *motion* condition was the one with the best performance. This finding confirms that *Physical motion cues* are more effective than *video-based cues* at influencing how people evaluate an autonomous robot that shows such cues. Our participants' rating of the cleanliness of the rooms further confirm such result: in the *motion* condition, the carpet's cleanliness was rated higher than in the *video* condition, after the robot's operations.

The qualitative data about why participants selected one specific robot like the one performing better provide further evidence of the effect of the two different notifications and their potential to influence people's perception of the robots. As illustrated in Figure 3.6, participants provided generic answers to this question only in three instances. In contrast, in the majority of cases, our participants' answers included specific and tangible reasons to support their choice, despite the fact that the Roombas did not actually clean either of the two rooms.

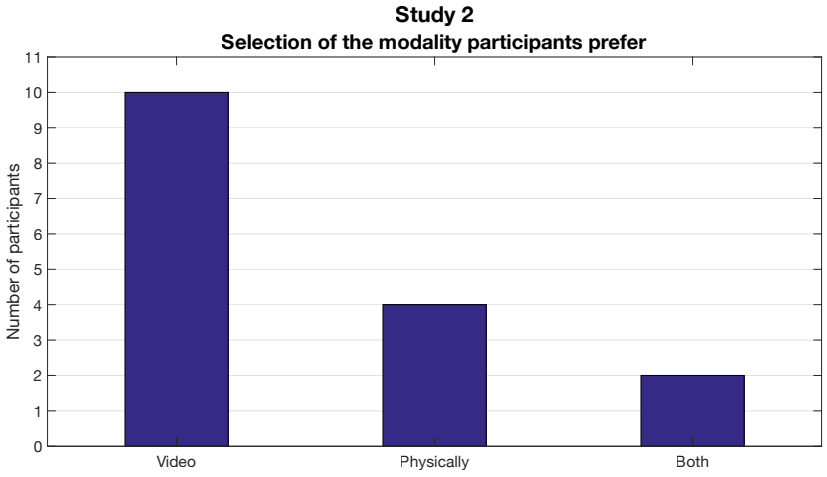

Figure 3.8: Selection of the participants for preferring one modality over the other in Study 1. Number of participants on the y-axis.

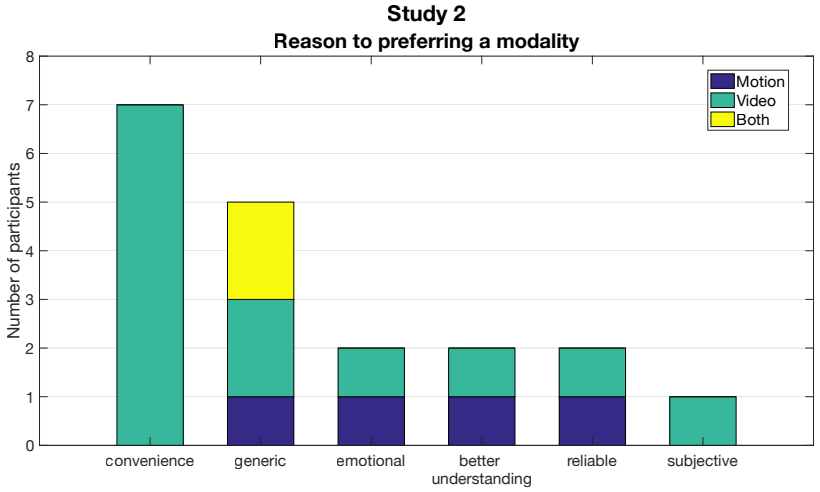

Figure 3.9: Reasons expressed by participants for preferring one notification modality over the other in Study 2.

Even though the performance ratings clearly indicate the *motion* condition as the most successful one, when participants were asked about their general preference regarding the notification style most of them chose the video. These results are an extension of Takayama et al. (2011) findings, they found that people enjoy more a virtual robot than a physical. The most frequent reasons to support this choice was convenience. Such contrast between performance ratings and general preference seems to suggest that participants were not aware of the bias that the *motion* condition caused on their performance rating. It should also be noted that while the performance rating was related to a financial incentive, the question about general preference was not. Therefore it is also possible that participants answered the latter more casually.

3.3 Study 3 – video-based cues vs. no-motion

From the first and second study, we learned that robot motion has an impact on peoples perception when experienced in person. However, the findings of Study 2 opened a new research question, Can *video-based cues* influence people's perception of autonomous robot's performance? To answer this question, we designed a third study to compare whether *video* feed has an impact on people's perception.

3.3.1 Method

3.3.1.1 Study Design

The design of Study 3 is a combination of conditions from Study 1 and 2. We compared *video* condition vs. *no-motion* condition. With this new configuration of conditions,

participants received text-based notifications and *video-based* notifications.

3.3.1.2 Participants

A total of 16 participants (6 male,10 female) took part in the study, and all of them were members of the University: undergraduate and postgraduate students, including a wide range, from Computer Science to Medicine, to Mechanics, and Psychology. The ages of the participants ranged from 22 to 32 years old ($M = 26.31, SD = 2.91$).

3.3.1.3 Equipment

The equipment used in this study was similar to Study 2; participants received the *video-based* notification and text-based notification on the web page we already design to be shown on the 13" laptop.

3.3.1.4 Procedure

For this study, we follow the same procedure as Study 2. The first part of the study was similar to Study 2. However, participants after the waiting time half of them received first the *video-based* notification, vice versa for the other half. Furthermore, the rooms and the robots were also alternated: half of the participants saw the robot in the *video* condition in Room A, and the other half saw the robot for the same condition in Room B. Moreover, in addition to the *reward-based* and *non-reward-based* questions, participants were asked whether they would like to receive *video-based* notifications and why.

3.3.2 Results

3.3.2.1 Selection of Roomba robot with the best performance.

In total 6 of the 16 participants considered that the Roomba in the *video* condition was the best one. Other 6 participants preferred the *no-motion* condition. The remaining 4 participants stated that both robots had the same performance. For the *non-reward-based* question, only one participant changed his answer from *no-motion* condition to both robots. These results are illustrated in Figure 3.10.

3.3.2.2 Reasons for choosing one Roomba over the other.

The responses to the question about why participants selected a particular Roomba as the one performing the best were summarised through thematic analysis. Each response

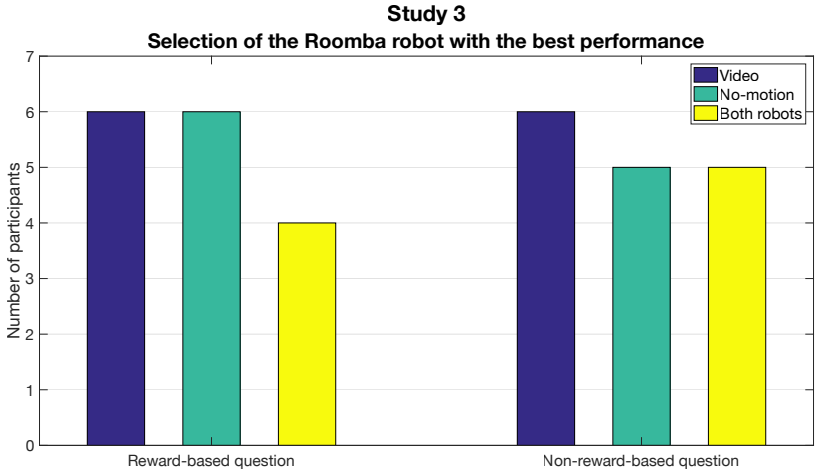

Figure 3.10: Selection of the participants for preferring a Roomba for the *reward-based* question and *non-reward-based* question in Study 3. Number of participants on the y-axis.

was associated with one or two themes, with six themes used in total: *details, relative, generic, room features, similar performance* and *clean already.* Figure 3.11 illustrates the frequencies of these themes for those who preferred the *no-motion* condition or *video* condition and those who preferred both conditions. The theme *details* was associated with responses which referred to specific issues in the room, such as "[...] cleaned the three particles I saw during the first inspection." and "there was a piece of hair that was not taken away by Roomba." The theme *relative* was used when the responses referred to the comparison of how clean the room was before and after the operation of the robots, such as: "The condition in room one has been improved." and "Did not notice any improvement in Roomba B [...]". Comments categorised as *generic* included "The room seem cleaner than with the other one." and "I felt the room was cleaner." The theme *room features* was used when participants referred to the influence of room features on the performances of the robots, such as "Because the cables in room A are lying on the floor." The theme *similar performance* was used when participants comment that both conditions have the same performance, such as "Both had the same performance. I did not notice a difference between them." and "I did the checking physically by touching and hitting the carpet just to see if there is dust floating around in the air. Surprisingly, both Roombas performed equally good." Finally, one participant stated that the room was clean to start with ("there was not much to clean"), so this response was categorised as *clean already.*

3.3.2.3 Evaluation of the Cleanliness of the carpets.

A Mann-Whitney Test revealed no statistically significant effect of no-motion and video on the rating of how clean the rooms were after the robots' operations. Figure 3.12

shows participant's evaluation of the cleanliness of carpets. No statistically significant differences were found in the ratings of how clean the rooms were before the robots' operations between the conditions.

3.3.2.4 Video-based notification preference.

Fifteen participants stated that they would like to have a *video-based* notification. However, one participant suggested that a text-based notification is enough as a feedback.

3.3.2.5 Reason for preferring a video-based notification.

The responses to the question about why they would like a *video-based* notification as a feedback were summarised through different themes. Each response was associated with one theme, with three themes used in total: *convenience, emotional,* and *reliance.* Figure 3.13 illustrates the frequencies of these themes. Comments categorised as *convenience* included "It is easier for people to notice the work has been done." and "I could see how the Roomba was working". Only one participant claimed that text-based notification is enough "the video is not necessary, the text message would be enough for me.", This comment was categorised in the same code as previous one. The theme *emotional* included "Looks more interesting and made the product look great quality." and "I liked it was live." Finally, the *reliance* code included "Give a feeling that the robot is working!!!" and "I believe the Roomba really finished cleaning".

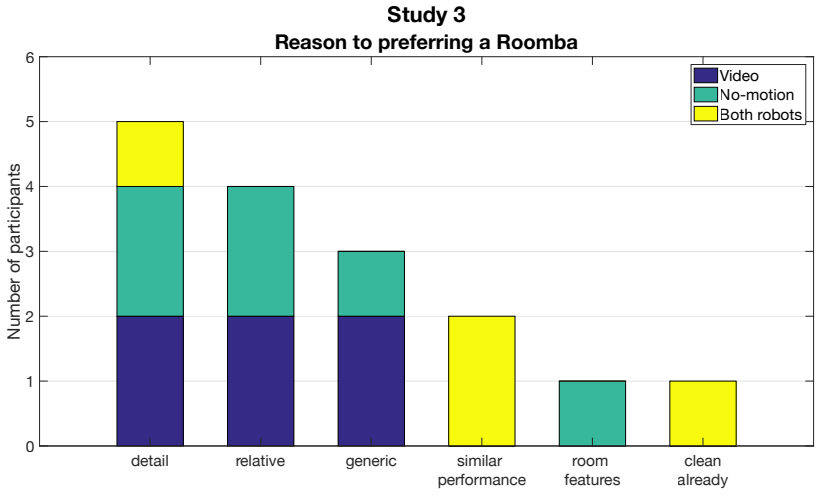

Figure 3.11: Reasons expressed by participants for preferring one Roomba over the other in Study 3.

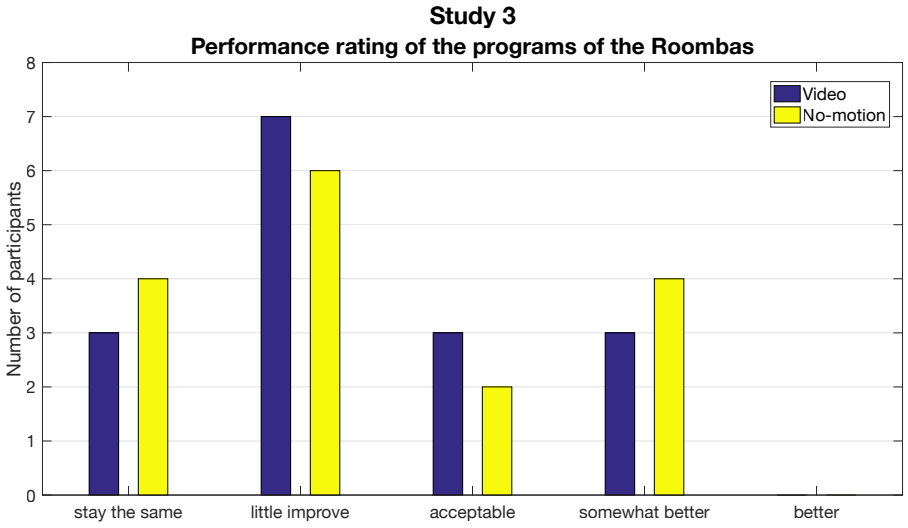

Figure 3.12: Likert-scale of participants' evaluation of the cleanliness of carpets for the *video* and *no-motion* conditions. Number of participants on the y-axis.

3.3.3 Discussion

The results of Study 3 did not yield a significant difference between the video and the *no-motion* condition. Participants' opinions were divided, almost exactly evenly, across the three options: some considered the robot that they saw moving in the video as performing better, some considered the robot they did not see moving as performing better, while others contemplated the performance of the two robots to be the same. These results suggest that *video-based cues* do not have a significant effect on people's perception. Furthermore, the qualitative findings suggest that in this study some participants noticed that the robots did not clean the carpets, as they stated that the rooms' cleanliness stayed the same after Roombas' operation.

Even though the video notification did not influence participants' perception of the robots' performance in a clear way, this notification modality was almost unanimously selected as the preferred one. This preference was justified through a variety of reasons, as summarised in Figure 3.13. These results suggest that *video-based* notifications for autonomous systems may have some advantages, but does not lead users to perceive the system to perform better.

3.4 Summary

In this chapter, we found that *physical motion cues* can affect how people perceive autonomous robots' performance. In more detail, Study 1 (N=16) results have confirmed

this finding by how participants rated higher autonomous robot's performance in comparison to a robot that did not show any physical motion feedback. These results extend what was previously reported by Hoffman and Vanunu (2013).

Additionally, Study 2 (N=16) findings support that *physical motion cues* influence people's perception of autonomous robots, compared to *video-based cues*. In more detail, *physical motion* is needed to see and effect on how people perceive system's performance, and that *video-based cues* instead is not enough. However, we found that people prefer to have a video-based notification rather than see the robot physically. However, the implementation of a video-based notification requires additional implementation of a webcam that can track robot's movement. In contrast, people's evaluation was higher for the robot that showed *physical motion cue*. These results suggest that participants were not aware of the bias that the *physical motion cues* caused on their performance rating.

Finally, in Study 3 (N=16), we compared *video-based cues* against a robot that did not show movement. These results suggest both conditions have the same effect on how people perceive autonomous robots' performance. Additionally, we observed that people still prefer to have a video-based notification as a feedback. In summary, the results of the three studies suggest that *physical motion cues* as a feedback affect how people perceive autonomous systems' performance. In particular, seeing the robot moving as it finishes its operation in person led our participants to rate its performance higher than not seeing any motion, or seeing the same motion over video.

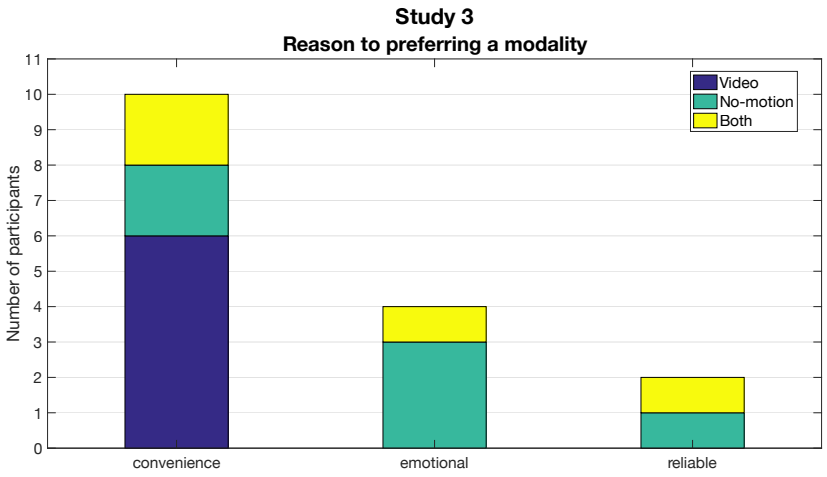

Figure 3.13: Reasons expressed by participants for preferring a *video-based* notification in Study 3.

Chapter 4

How Animation Cues Change People's Perception of HWR and POS Screen-based Systems' Performance

As we observed that motion could influence the perception of the performance of robots, we set out to assess the potential for *animation cues* to affect users' perception of the performance of smart systems. We consider this investigation as an attempt to generalise the results found in Chapter 3, where we observed that motion could influence the perception of the performance of vacuum cleaning robots. As an extension of these findings, we designed three studies to address our the fourth research question "Can *animation cues* influence people's perception to score higher smart system's performance?". Hence, Study 4 (N=16) and Study 6 (N=16) were designed to analyse whether *animation cues* also has a positive effect on people's perception of the performance of handwriting recognition and part-of-speech systems respectively. Moreover, Study 5 (N=16) and Study 6 were designed to test this effect in a non-controlled environment to analyse whether the effect can be observed outside a lab environment. In more detail, we ran our Studies 5 and 6 in the crowdsourcing platform Amazon Mechanical Turk.

4.1 Study 4 – Animation cues vs. no-animation (HWR system), Lab Study

We designed a lab study involving a relatively simple graphic animation, somewhat related to the system operation. We chose an HWR system, a system that recognises handwritten text and converts it to electronic text (or e-text / typed text), because this

is a common task that many people can relate to, at least conceptually, and it also can be simulated easily (Verame et al. (2016)). Moreover, we chose to use text in Filipino, a language that most users would be unlikely to know, to mimic the likely circumstances of casual users not being familiar with the kind of data handled by the system. In this way, rather simply checking the system output for typos, users are required to compare the input and the output looking for differences, a task that is more attention demanding.

4.1.1 Method

4.1.1.1 Study Design

A fully counterbalanced, a within-participants design was used, where participants were asked to evaluate and compare the performance of two HWR systems, each corresponding to an experimental condition: *animation* and *no-animation*. Both systems were based on the same graphical user interface, illustrated in Figure 4.1. On the left-hand side of the screen, a scan of a page of handwritten text in Filipino (system's input) is displayed, while on the right-hand side the typed text (system's output) is shown. In both cases, the interface screen was preceded by a 'loading' screen, showing just a text that the system was processing its task for 10 seconds, reinforcing the idea that the system was doing something in the background. In the *no-animation* condition, the system presented the result immediately after the loading screen, and *no motion cues* were displayed. In the *animation* condition, after the loading screen, an animation was shown: on the last two lines of the input words were highlighted one by one, with a delay of a few hundred milliseconds; as each handwritten word was highlighted, the corresponding word on the output appeared. The first word highlighted by the animation was the word "naging", for more detail see Figure 4.1. By highlighting the words one by one, our intention was to give users an impression of how an algorithm may process the input data[1].

External validity was a key factor in this study. Therefore both HWR systems showed the same handwritten text, and both systems involved the same number of errors (four mistakes per paragraph, resulting in a total of eight mistakes across two paragraphs). In the last two sentences (the ones highlighted by the animation) both systems presented one error. Additionally, a consensus-oriented reward mechanism was adopted to try and ensure that participants would provide a meaningful and thoughtful evaluation when they choose which system they considered to have the best performance. Participants were told that if they select the system which the majority identified as the one with the best performance, they will be rewarded with a £10 voucher at the end of the experiment. This question is later referred to as the *reward-based question*.

[1] https://vimeo.com/183480644.

4.1.1.2 Participants

A total of 16 participants (10 male, 6 female) took part in the study, and all of them were students: undergraduate and postgraduate, from a wide range of backgrounds, from Computer Science to Mechanics, Psychology and Design. Participants were recruited through adverts posted on university social network groups. The age of participants ranged from 18 to 24 years old (M= 20.68, SD=1.70).

4.1.1.3 Equipment

The study was run in a room at a university, where each participant sat with the investigator. The interfaces and the questionnaire were implemented as a simple Web application, using HTML5 and Python with the Django framework. The application was displayed on a 13" laptop and served from the same computer each time. The animation was a GIF image.

4.1.1.4 Procedure

At the beginning of the study, participants received written instructions asking them to evaluate and compare the performance of the two HWR systems. The two systems were presented one at a time, in sequence: half of the participants first experienced the *animation* condition, while the other half first experienced the *no-animation* condition. In each condition the system was shown to participants for two minutes, so they had a limited time to compare input and output. After the participants had seen both systems, they were asked to fill in a questionnaire to evaluate their performance. Participants were first asked to rate the individual performance of each system on a 5-point Likert scale. They were then asked to select which of the systems they believe the majority of

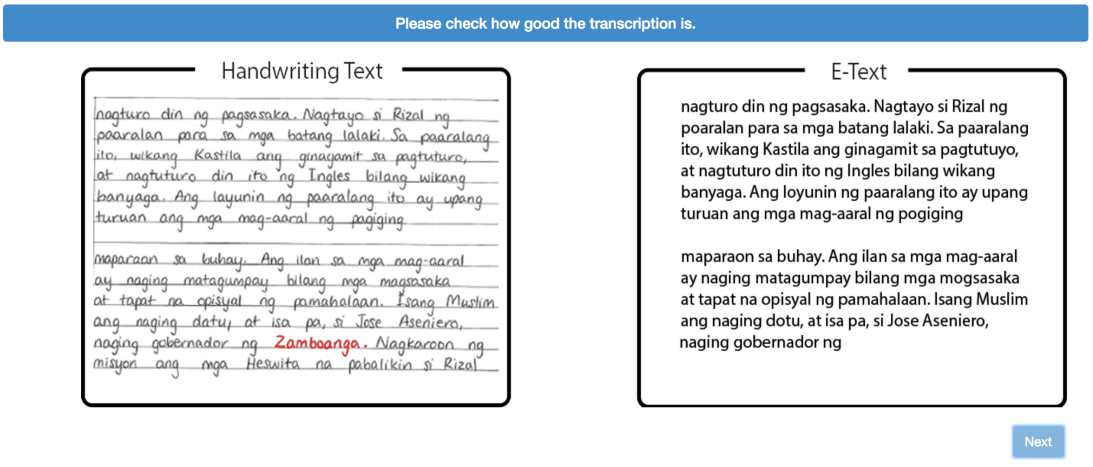

Figure 4.1: Interface of the HWR system implemented in Study 4.

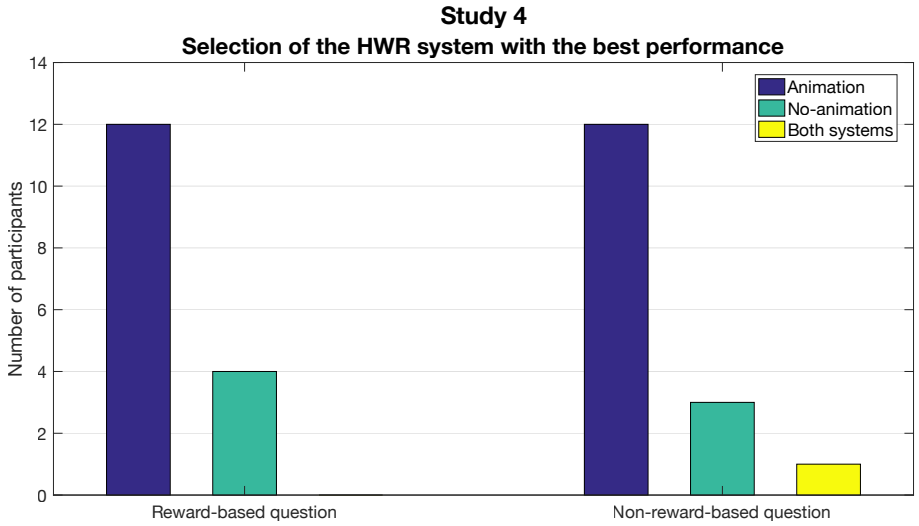

Figure 4.2: Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 4. Number of participants on the y-axis.

participants would choose to have the best performance and to provide a justification for their selection. As mentioned above, if participants answered in the majority they would receive a reward voucher. Additionally, we asked participants to describe in a text field why they chose one system over the other. After these questions were answered, participants were also asked (on a separate page) which system they considered to have the best performance without considering what the majority of the participants would choose, and no reward (we refer to this as the *non-reward-based question*).

4.1.2 Results

4.1.2.1 Selection of the system with the best performance

For the *reward-based* question, 12 of the 16 participants (75%) chose the system in the *animation* condition as the one with the best performance, the remaining 4 participants (25%) chose instead the one in the *no-animation* condition, while nobody indicated that both systems had the same performance level. Only 1 participant answered the *non-reward-based* question differently than the *reward-based* one, changing the answer from "no-animation" to "both systems". These results are illustrated in Figure 4.2.

4.1.2.2 Reasons for choosing one system over the other

Participants' responses to the question about why they selected a particular HWR system as the one that performed best were categorised through thematic analysis. Each

response was associated with one or two of the following three themes: *number of errors, type of errors,* and *generic*. Figure 4.3 illustrates the frequencies of these themes. The theme *number of errors* was associated with responses where the participants reported finding fewer errors or mistakes in the output of one system than in the output of the other, such as "There were fewer mistakes in total", "It has mistakened less characters." and "I think both of them had about the same number of errors, however the second one's were more obvious [..]" Comments categorised as *type of error* were linked to situations when participants pointed out typographical errors they found, such as "only confuses a-o, b-h, ri-n whereas the second also confuses d-g" and "[..] algorithm only got mistakes when the words contain 'a' and 'o'." Finally, comments such as "More sensitive recognition of lettering [...]", and "[...] Errors of the second program are easier to guess and find out." were categorised as *generic*.

4.1.2.3 Performance ratings

A Wilcoxon Signed-rank Test revealed that the performance evaluation was higher for the *animation* condition ($Mdn = 4$) than for the *no-animation* condition ($Mdn = 3.5$), ($Z = 2.07, p < .05, r = 0.37$). Figure 4.4 shows participant's evaluation of the performance of the systems.

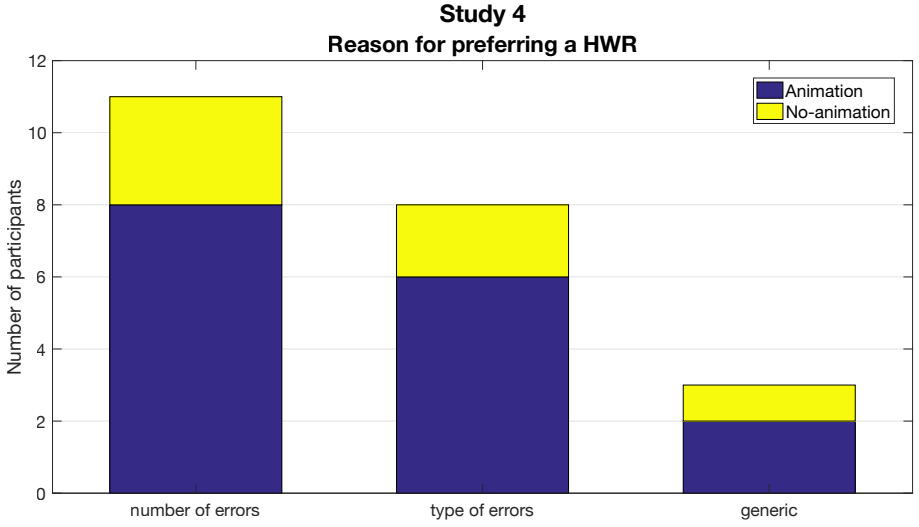

Figure 4.3: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition in Study 4. Number of participants on the y-axis.

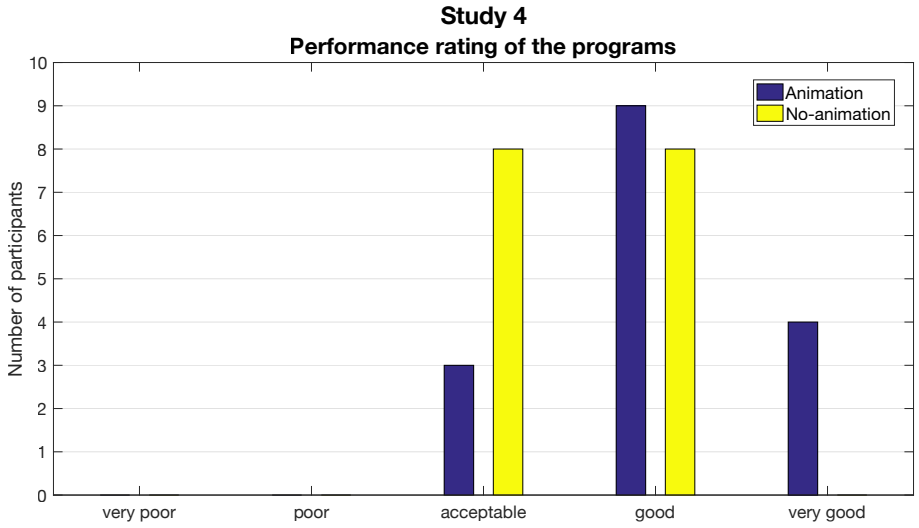

Figure 4.4: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

4.1.3 Discussion

The results of Study 4 show that *animation cues* have an effect on participants' perception of the system's performance. The data shows clearly that the majority of participants considered the performance of the system in the *animation* condition to be better. It should be noted that this was the case despite the fact that one error was present in the sentence highlighted by the animation. In other words, even though the animation could have drawn the participants' attention to the mistake, for most of them the animation instead had the opposite effect. The qualitative data further supports this result; most participants seem to believe that the system in the *animation* condition made fewer errors or different kind of errors than the other system, despite the two systems producing the same number and kind of errors. Moreover, participants seemed to be unconscious of the effect: none of the comments referred explicitly to the animation.

These results extend those from Chapter 3, who showed that *physical motion cues* could be used to change people's perception of the performance of vacuum cleaning robots. Our results indicate that the effect of motion does not apply only to physically moving systems, but also to graphical user interfaces through animation.

These results open up a number of follow-up questions. Can this effect be observed in a less controlled environment? Can it be observed for a different type of smart system? The following two studies were designed and carried out to address these two questions.

4.2 Study 5 – Animation cues vs. no-animation (HWR system), MTurk study

To assess whether similar results to those of Study 4 can also be observed in a less controlled environment than the lab, we decided to run a similar study on a crowdsourcing platform: Amazon Mechanical Turk [2] (MTurk). In addition to being less controlled, crowdsourcing environments are also reported to include more diverse participants (Buhrmester et al. (2011); Germine et al. (2012)). Crowdsourcing has become a widespread online tool that researchers and companies use to outsource micro-tasks. In MTurk, these micro-tasks are referred to as Human Intelligence Tasks (HITs)[3], that leverage human computation, gather distributed and unbiased data or validate results (Difallah et al. (2015); Kazai et al. (2013); Mason and Watts (2010)). The people who complete such crowdsourcing tasks are referred to as 'crowd workers,' or simply 'workers.'

4.2.1 Method

4.2.1.1 Study Design

The study design was the same as in Study 4: fully counterbalanced, within-participants, where participants were asked to evaluate and compare the performance of two HWR systems, each corresponding to an experimental condition: *animation* and *no-animation*. The two conditions were identical to Study 4. However, we added an extra question in the post-task questionnaire asking participants to justify their selection for the non-reward-based question.

Because of the constraints of the MTurk platform, the reward mechanism was adjusted accordingly. Participants received a *fixed reward*, to compensate them for the time they spent working on our study, as well as an additional *performance-based, consensus-oriented reward*, designed to increase the ecological validity of the study and motivate the participants to provide thoughtful responses, similar to Study 4. To make sure that participants received a fair payment, we considered the minimum wage across the different countries participants could be from (see restrictions below), and we selected the Canadian one as the one currently highest, at approximately $10 per hour. Therefore, the fixed reward was set to $1.17 – the whole task takes about 7 minutes, $1.17 corresponds to about 10 minutes, leaving some margin. This amount was awarded as soon as all participants completed the study. The performance-based, consensus-oriented reward was awarded as a "bonus" on the MTurk platform, and it amounted to the same value as the fixed reward. In other words, the performance-based, consensus-oriented reward

[2] https://www.mturk.com.
[3] https://www.mturk.com/mturk/help?helpPage=overview.

doubled the money that participants received for the study. It was awarded once all participants had completed the study.

4.2.1.2 Participants

Participants were recruited through MTurk, with two restrictions. First, they were only allowed to take part in the study if their location was *United States, Canada, or Australia*, to avoid issues related to English comprehension. Second, recruitment was limited to participants with HIT approval rate was equal to 100% (this is the approval from those who advertise the HITs[4]), as a rejection on MTurk often indicates that workers do not take tasks seriously. A sample of 16 participants successfully completed the online study, to keep the same pool size as Study 1 so that we were able to compare the two studies.

The age of participants ranged from 21 to 44 ($M = 33, SD = 5.94$), with 12 of them being males (75%) and 4 being females (25%). All our participants were United States nationals. The education levels of the participants ranged from secondary school level to master's degree or (equivalent). Overall 10 of them had a university degree level, 5 a secondary school level and 1 master's degree level. None of our participants reported knowing Filipino.

4.2.1.3 Equipment

We used the same Web application developed for Study 4. However, this was extended with an initial questionnaire to obtain the participant's demographic information, and it was deployed to a publicly accessible Web server, to allow MTurk' workers to access it from their personal computers. On the MTurk platform, we added the URL link in the HIT where workers could access it.

4.2.1.4 Procedure

Before participants accepted the task, they were told that the aim of the experiment was to compare two different HWR systems, one at a time[5]. They were instructed to check the system's outcome and find possible mistakes that the system could make in the transcription of the handwritten text to typed text. After the introduction, the participants can decide to either accept or reject the task. Once they decided to accept the task, an external link was displayed. The link opened a new window that showed a brief questionnaire, asking for the participants' demographic information.

[4] https://www.turkprime.com/Home/FrequentlyAskedQuestions.
[5] This information was displayed in the HITs' description.

After the participants had answered the initial questionnaire, the study followed the same procedure as Study 4, with the addition of an extra question in the post-task questionnaire (asking participants to justify their selection for the non-reward-based question), as mentioned above. At the end of the task and once we approved their participation in our study, we flagged the participants by awarding an 'MTurk qualification' to ensure they are unable to take part in our follow-up studies.

4.2.2 Results

4.2.2.1 Selection of the system with the best performance

For the *reward-based* question, 12 of the 16 participants (75%) chose the system in the *animation* condition as the one with the best performance, 3 participants (19%) chose the system instead of the *no-animation* condition, while the remaining one indicated that both systems had the same performance level. Only 1 participant answered the *non-reward-based* question differently than the *reward-based* one, changing the answer from "animation" to "no-animation". These results are illustrated in Figure 4.5.

4.2.2.2 Reasons for choosing one system over the other – reward-based question

The responses to the question about why participants selected a particular HWR as the one the majority would choose to have the best performance were summarised through

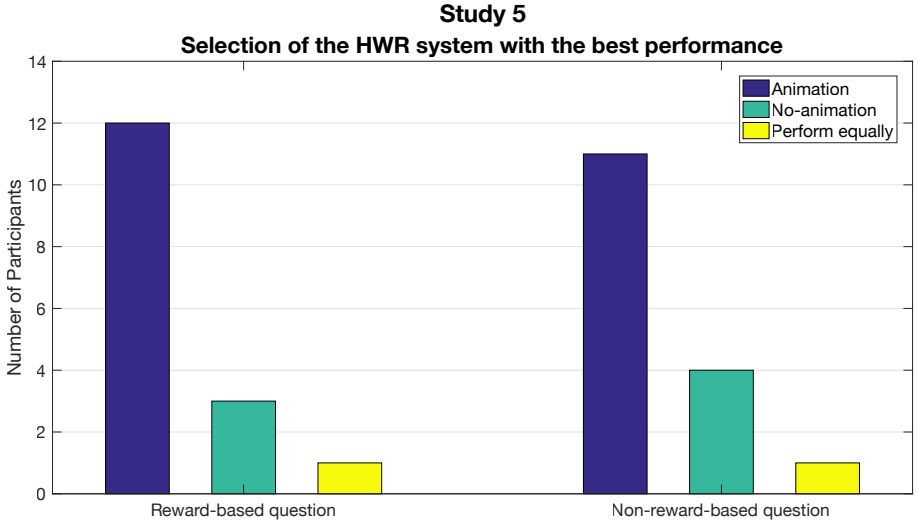

Figure 4.5: Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 5. Number of participants on the y-axis.

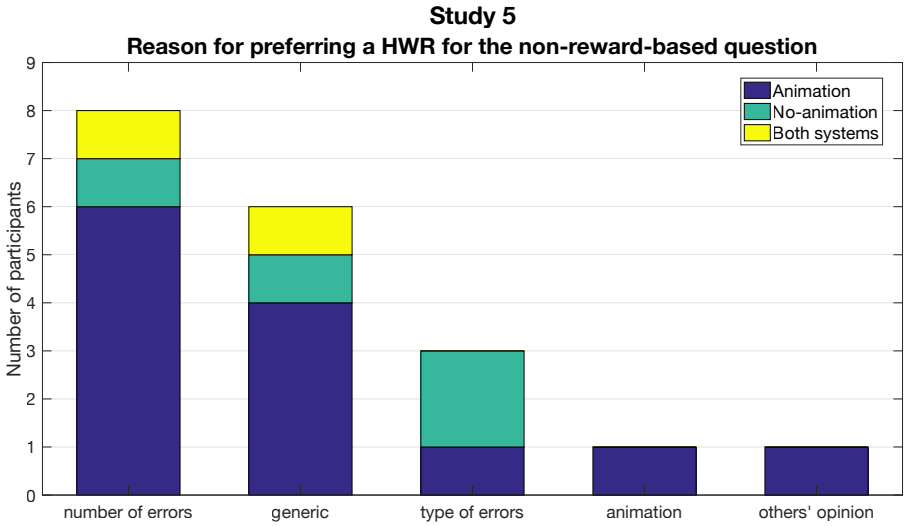

Figure 4.6: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

thematic analysis. Each response was associated with one or two of the following five themes: *number of errors, generic, type of error, animation,* and *others' opinion.* Figure 4.6 illustrates the frequencies of these themes. The categories *number of errors, generic,* and *type of errors* were the same as Study 1. The category *animation* was used when comments were related to the animation, e.g.: "Actually seeing the words transcribed probably leaves a good impression." Finally, the response of one participant that selected the HWR of the *animation* condition based on what other participants would select ("I think the second works better because the workers are more prepared at that point.") was categorised as *others' opinion.*

4.2.2.3 Reasons for choosing one system over the other – non-reward-based question

Thematic analysis was also applied to the answers related to the non-reward based question. The same themes as the previous question emerged, their frequencies are reported in Figure 4.7. For the majority of participants, 13 out of 16, the answers were the same (in terms of themes) as for the reward-based question. Only 3 participants answered this question differently than the previous one, and of these 3 only 1 changed their selection. In particular, the participant who selected a different system commented that they believed that other participants would choose the system in the *animation* condition because of the animation itself ("Actually seeing the words transcribed probably leaves a good impression."). Regarding the other 2 participants who provided different reasons without changing their selection, in one case the answer went from "type of errors and number of errors" to just "number of errors", on the other case, it went from "number

of errors" to "animation" ("I liked that the second computer program highlighted the text in red as it was transcribing it").

4.2.2.4 Performance ratings

A Wilcoxon Signed-rank Test revealed that the performance evaluation was higher in the *animation* condition ($Mdn = 5$) than in the *no-animation* condition ($Mdn = 4$), ($Z = 2.45, p < .05, r = 0.43$). Figure 4.8 shows participant's evaluation of the performance of the systems.

4.2.3 Discussion

The results of this study clearly showed that the positive effect of *animation cues* persist even in a less controlled environment. The majority of participants reported that the system which integrated the animation performed better than the other system, for both the reward-based and non-reward-based questions. The statistically significant differences in the Likert scales results, as well as the qualitative data further confirm this finding. Moreover, similar to those from Study 4, the level of detail of the responses we collected clearly shows that the participants were committed to the task, giving credibility to the data. For example, participants referred not only to the number of errors that they found in the transcribed text but also to the type of errors (e.g., "I think the first program had more problems distinguishing the 'a' and 'o'.") Only one participant answered differently between the reward-based and the non-reward-based

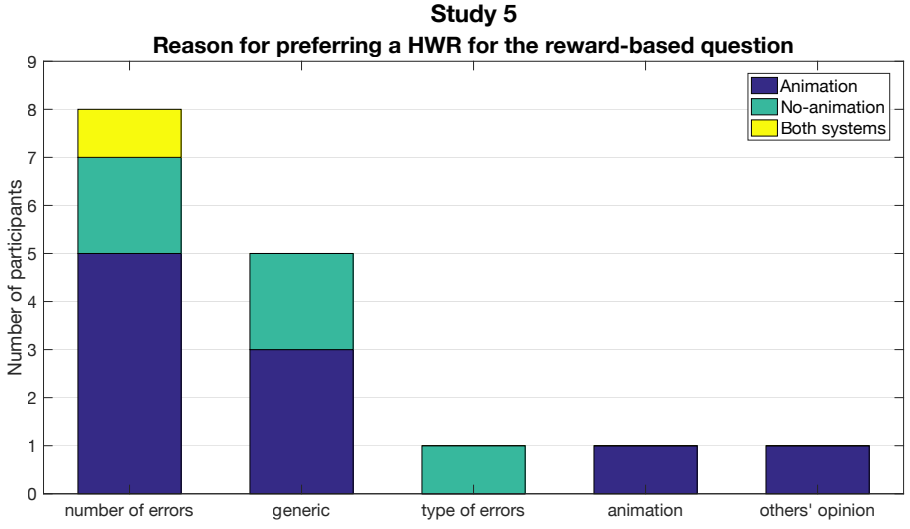

Figure 4.7: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the non-reward-based question. Number of participants on the y-axis.

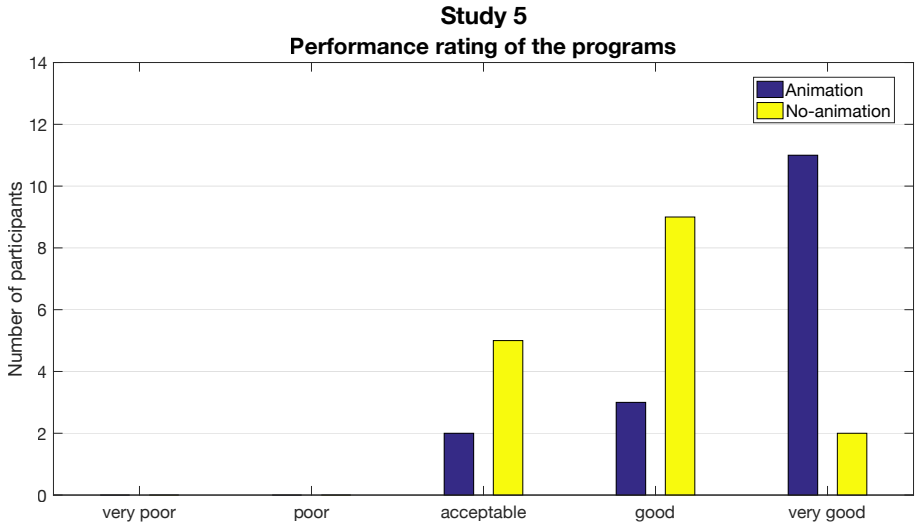

Figure 4.8: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

questions. This finding suggests that the financial incentives did not solely motivate answers.

Only two participants justified their selection in terms of others' opinions or impressions, and by explicitly referring to the animation. This finding can be interpreted as confirming that the effect of *motion cues* is mostly unconscious. It should also be noted that these two participants did not change their answer between the reward-based and non-reward-based questions. This appears counterintuitive, and it can perhaps be explained as these two participants not paying much attention to the non-reward-based question. It should be noted that these references to others' opinions and the animation emerged in Study 5, but not in Study 4. Such difference can be explained by the less controlled nature of Study 5.

The alignment of the results from Studies 4 and 5 also indicates that to investigate this phenomenon further follow-up studies can be conducted on the MTurk platform, with considerable practical advantages. Having observed the effect of *animation cues* in a less controlled crowdsourcing environment, we turn to investigating whether this effect is specific to the handwriting recognition system we used so far, or whether the results can be generalised to a different type of smart system.

4.3 Study 6 – Animation cues vs. no-animation (POS system)

Studies 4 and 5 tested the effect of *animation cues* using one particular system, a HWR system. Handwriting recognition is by its very nature a visual task, making us wonder whether this factor alone may explain our results. So we designed a new study to assess whether the same effect would occur with a different type of system, and one involving processing that is not visual in nature. We selected a part-of-speech (POS) tagging system, a system that analyses natural language sentences and tags each word according to its syntactic function, such as article, adjective, adverb, conjunctions, noun, preposition, pronoun, and verb. POS tagging algorithms are readily available through open source libraries[6], and their application has been suggested for different types of interfaces and visualisations (Chuang et al. (2012); Yatani et al. (2011)). We decided to continue to use text as the type of data handled by the smart system, for continuity with the previous studies and hence facilitate comparison of the results.

4.3.1 Method

4.3.1.1 Study Design

The study design was almost identical to Study 5: fully counterbalanced, within-participants, where participants were asked to evaluate and compare the performance of two systems,

[6] http://www.nltk.org/.

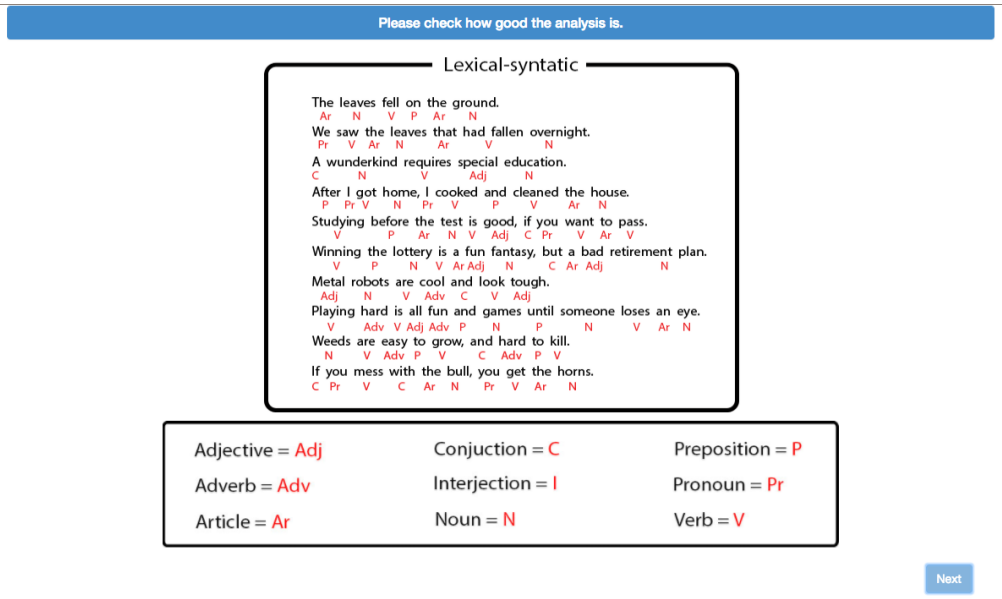

Figure 4.9: User Interface of the Part-of-Speech tagging system implemented in the Study 6

each corresponding to an experimental condition: *animation* and *no-animation*. There were only two differences. First, the two conditions were applied to a POS tagging system on a piece of text in English, rather than an HWR one (in Filipino), as illustrated in Figure 4.9. Specifically, in this study, the animation shows the tags of the two last sentences appearing a few milliseconds one after the other next to each of the corresponding words[7]. Second, interpreting the results of the POS tagging system requires familiarity with POS tagging as a grammatical exercise. Because not everyone might be familiar with this, we included a *validation task*: participants had to tag a given sentence (in English) with the POS corresponding to each word. Only those who completed this validation task with less than 3 mistakes (out of 8 words) were allowed to proceed to the main task. Such validation task was not necessary for the HWR system because anyone could complete that one by visual inspection.

The reward structure was also identical to Study 5, with a fixed amount being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority. To account for the slightly longer duration of the study (1 extra minute, due to the validation task) the reward for Study 6 was $1.33.

4.3.1.2 Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 5. The sample size was again 16, reported age ranged from 21 to 54 ($M = 26, SD = 10.02$), 5 males (31%) and 11 females (69%). 12 of them were United States nationals, 1 South Korean, 1 Canadian, 1 Belgian and 1 Bangladeshi national. The education levels of the participants ranged from primary school level to doctoral degree level or equivalent. Overall, 9 participants had a university degree, 4 completed secondary school, and 3 completed primary school.

4.3.1.3 Equipment

The Web application used for Studies 4 and 5 was modified to include the validation task described above, and to show the POS tagging system in place of the HWR one. As in Study 5, the Web application was deployed to a publicly accessible Web server, to allow MTurk' workers to access it from their personal computers.

4.3.1.4 Procedure

This study followed the same procedure as Study 5, with the exception of the additional validation task outlined above. The validation task was displayed after the initial questionnaire about demographic information and before the main task. In the validation,

[7] https://vimeo.com/210299892.

task participants were shown a POS-tagged sentence as an example. Then they were asked to tag one sentence. As mentioned above, if participants made less than 3 mistakes in this exercise they proceeded to the comparison of the two experimental systems, as in Study 5.

4.3.2 Results

4.3.2.1 Selection of system with the best performance

For the *reward-based* question, 11 of the 16 participants (69%) selected the system in the *animation* condition as the one with the best performance, 2 participants (12%) selected instead the system in the *no-animation* condition, while the remaining 3 (19%) expressed that both systems had the same performance. Moreover, one participant changed her choice for the *non-reward-based* question from "both systems" to "animation". The results are illustrated in Figure 4.10.

4.3.2.2 Reasons for choosing one system over the other – reward-based question

The responses to the question about why participants selected a particular POS as the one the majority will choose as the best performing system were summarised through thematic analysis. Each response was associated with one or two of the following three themes: *number of errors, generic* and *type of errors*. Figure 4.11 illustrates the frequencies of these themes. The themes were similar to those emerged from previous

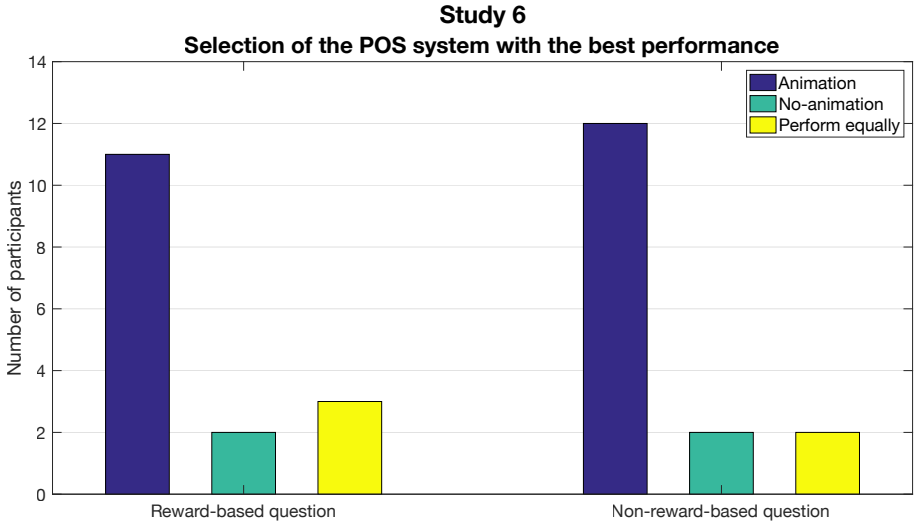

Figure 4.10: Selection of the participants for preferring a POS for the reward-based and non-reward-based questions in Study 6. Number of participants on the y-axis.

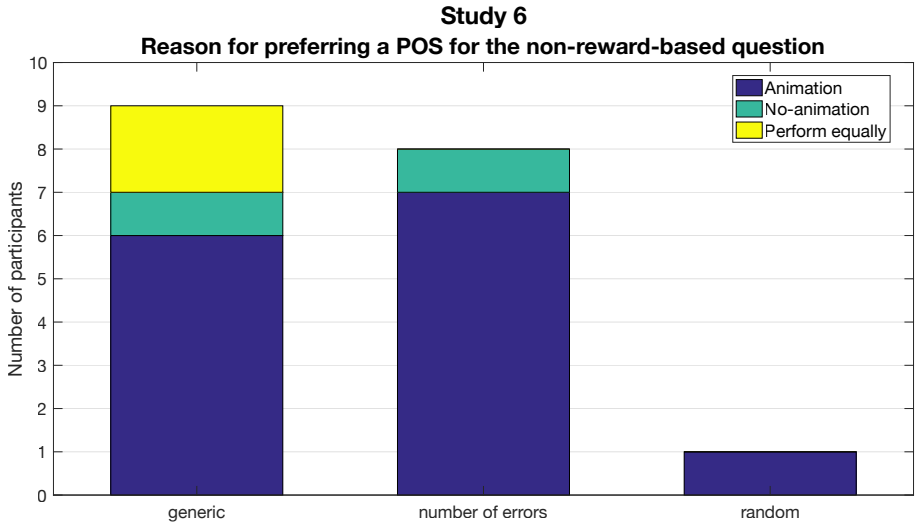

Figure 4.11: Reasons expressed by participants for preferring a POS in the *animation* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

studies, with the exception that "type of errors" referred to specific POS tagging errors ("It seemed to have less mistakes. For example, the first one called 'that' an article").

4.3.2.3 Reasons for choosing one system over the other – non-reward-based question

Participants' responses about why they personally considered a particular POS system as the best performing one or the systems had the same performance were summarised through thematic analysis. Each response was associated with one or more of the following three themes: *generic, number of errors,* and *random.* Figure 4.12 illustrates the frequencies of these themes. The themes were similar to those which emerged from previous studies.

For the majority of participants, 12 out of 16, the answers were the same (in terms of themes) as for the reward-based question. Only 4 participants answered this question differently than the previous one, and of these 4 only 1 changed their selection. In particular, the participant who selected a different system commented that she considered that both had the same performance, but she left a comment that her selection was random (e.g. "It's really a toss-up. I saw the same potential errors on the same word in both programs, so I'm just picking one."). Regarding the other 3 participants who provided different reasons without changing their selection, in one case the answer went from "type of error" to just "generic", in the second case went from "number of errors" to "number of errors and generic", in the last case, it went from "number of errors and type of errors" to just "number for errors".

4.3.2.4 Performance ratings

A Wilcoxon Signed-rank Test revealed that the performance evaluation was higher in the *animation* condition ($Mdn = 4$) than in the *no-animation* condition ($Mdn = 3$), ($Z = 2.55, p < 0.05, r = 0.45$). Figure 4.13 shows participant's evaluation of the performance of the systems.

4.3.3 Discussion

The results of Study 6 extend those of Study 5, they demonstrate that the effect of *animation cues* on participants' perception of system performance also applies to the POS tagging system we tested. The majority of participants selected the system in the *animation* condition as the one with the best performance, and the Likert-scale ratings for this system were higher, in aggregate than those for the *no-animation* condition, with statistical significance. Similar to Studies 4 and 6, the qualitative data collected in Study 6 indicates that participants offered a variety of reasons to justify their selections, and only 1 participant provided different answers based on the financial incentives, suggesting that most answers were not based solely on the financial incentives. Moreover, the themes emerged from the qualitative data are the same as Studies 4 and 5, further confirming the similarity of the effect on POS and HWR systems. Such effect, then, appears to apply even if the task performed by the system is not inherently visual, and hence if the animation does not directly mimic the task performed by the smart system.

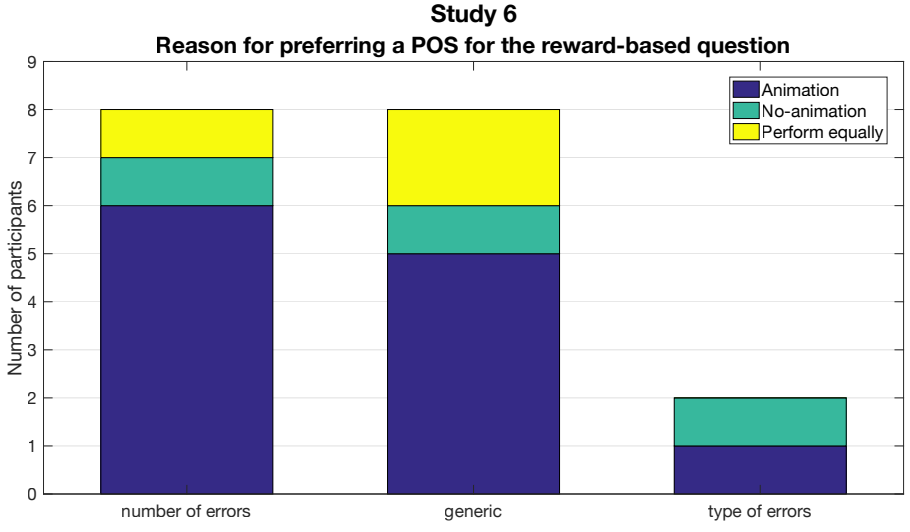

Figure 4.12: Reasons expressed by participants for preferring a POS in the *animation* or *no-animation* condition for the non-reward-based question. Number of participants on the y-axis.

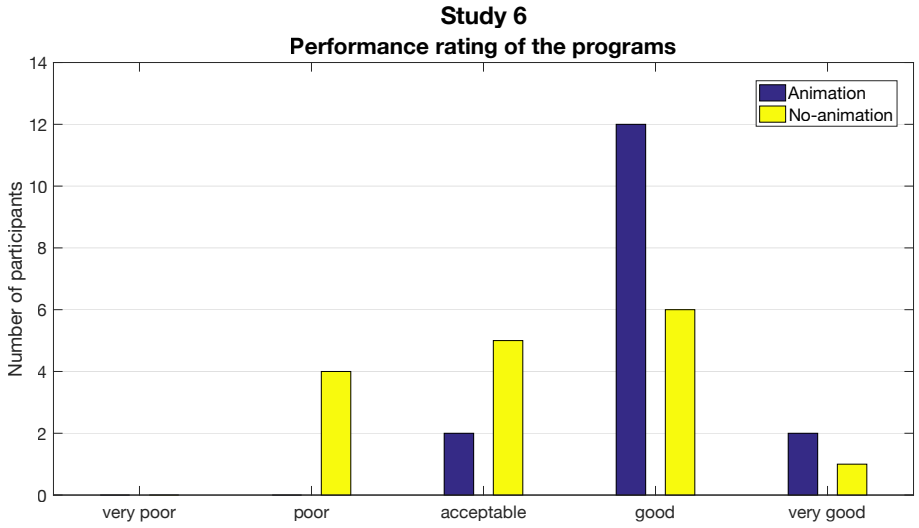

Figure 4.13: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

Having observed this effect both in the lab and on MTurk, and on two different systems, we turn to the question of *why* such effect occurs. Given that both animations highlight one word at a time, in reading order (from left to right), one option could be that the animations give users the impression that the systems process text in the same way a person would process it. Is it possible that the similarity to humans may positively influence users' attitude towards the system? This, in turn, may lead them to evaluate its performance more favourably, perhaps somehow suggesting to them that the system is "as smart as a person". An alternative explanation might involve more generally the relationship between the animations in our studies and users' mental model of how the system works (Balijepally et al. (2015)). In more detail, Balijepally et al. found that the quality of a developer's mental model positively impacted the users' performance as measured in terms of software quality. Moreover, the accuracy of a person's mental model of a system is based on the person's prediction of a system's future behaviour and therefore influences how they interact with it (Norman (2013)). The animations might induce a mental model that leads them to rate the system performance more positively. To assess the validity of these possible explanations, we designed and carried out three follow-up studies that we report in the following.

4.4 Summary

In this chapter, we focused on analysing whether the effect showed in Chapter 3 that *physical motion cues* have an effect on people's perception can be observed on screen-based systems. Thus, we designed three studies to analyse how *animation cues* have

an effect on people's perception of screen-based systems' performance. The findings of Study 4, 5, and 6 suggest that *animation cues* clearly have an effect on people's perception. In more detail, in Study 4 (N=16) we found that people score higher an HWR system that shows *animation cues* as a feedback than a system that did not show any feedback of how performs its task. Furthermore, the findings showed that people considered that the system made fewer mistakes that the system that did not show any *visual cue*.

In addition, the results of Study 5 (N=16) showed that the positive effect of *animation cues* persist even in a less controlled environment. Moreover, we found that the level of detail of the responses we collected from participants shows that the participants were committed to the task. Furthermore, the findings of Studies 4 and 5 indicated that this phenomenon, follow-up studies could be conducted in a less controlled environment (i.e. crowdsourcing platform MTurk). Moreover, the aim to use crowdsourcing platform was to have more diverse participants (Buhrmester et al. (2011); Germine et al. (2012)). Finally, we ran a Study 6 (N=16) to analyse whether the effect observed in the previous studies is particularly to the HWR system, or whether the results can be generalised to others screen-based systems.

The findings of Study 6 extend those of the previous two studies, they demonstrated that the effect of *animation cues* on people's perception of system performance apply to other screen-based systems. In more detail, we found that people score higher the performance of a part-of-speech system that shows *animation cues*. In summary, the findings of this chapter are an extension of the results we found on Chapter 3. As such, we can suggest that *visual cues* as feedback have an effect on how people perceive the performance of smart systems.

Chapter 5

What Makes Animation Cues Affect People's Perception?

As we observed in Study 4, 5, and 6, *animation cues* have a positive effect on people's perception. In comparison with Study 1 and 2, we observed that the *physical motion cues* also have a positive effect. However, the animation that we implemented in Study 4, 5, and 6 are not a functional part of system task. Thus, we need to understand which characteristics the *animation cues* require to assure the effect on people's perception persist when designers implement this *visual feedback* in their systems. For this reason, we present in this chapter three studies that we conducted on MTurk to find the reason behind why the *animation cues* affect people's perception. In more detail, Study 7 (N=16) was designed to address the research question "Do *animation cues* have an effect on people's perception of a smart system's performance if the animation of the system processing the task is similar to how a human would process the task?". Study 8 (N=16) was designed to find a relation between the animation we implemented in early studies and people's mental model of how they consider handwriting recognition systems perform their task. Finally, we designed Study 9 (N=64) to address the research question "Do *animation cues* have an effect on people's perception of a smart system's performance only if the animation is consistent with people's mental model of how the system works?"

5.1 Study 7 – NHL-animation cues vs. no-animation

Through Studies 4, 5 and 6 we found that *animation cues* can influence how people evaluate the performance of smart systems. Why do these *animations cues* have an effect on participants' perception of the smart system's performance? Could the effect be because the animations making the systems appear to process information like a human? If this is the case, then showing an animation where the order in which the

words are processed is decisively not human-like should have no effect on participants' perception of the system performance. So we designed a seventh study to test whether an animation that is decisively not human-like would still cause the same effect as the animation used in the previous studies.

Study 6 revealed that the animation effect applies in a similar way to a POS tagging system as it does to an HWR system. For simplicity, then, we decided to conduct further experiments on the HWR system, as it does not require the additional training and validation task described above.

5.1.1 Method

5.1.1.1 Study Design

The study design was identical to Study 5: fully counterbalanced, within-participants, where participants were asked to evaluate and compare the performance of two systems, each corresponding to an experimental condition: *non-human-like animation* (*NHL-animation*, for short) and *no-animation*. The only difference was the animation: instead of highlighting words in left-to-right order (as in Studies 4, 5 and 6), the order was random[1].

The reward structure and amounts were also identical to Study 5, with a fixed amount of $ 1.17 being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority.

5.1.1.2 Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 5. The sample size was again 16, age ranged from 22 to 44 ($M = 31, SD = 7.16$), 13 males (81%) and 3 females (19%). All except for one of the participants reported being United States nationals, and the remaining one German. The education levels of the participants ranged from primary school level to university degree level or equivalent. Overall, 7 participants had a university degree, 7 completed secondary school, and 2 completed primary school. None of the participants reported knowing Filipino.

5.1.1.3 Equipment

The same Web application used for Studies 4 and 5 was used in Study 7, with the only difference of the animation, as described above.

[1] https://vimeo.com/183550733.

5.1.1.4 Procedure

The procedure was the same as Study 5.

5.1.2 Results

5.1.2.1 Selection of the system with the best performance

For the *reward-based* question, 10 of the 16 participants (62%) selected the system in the *NHL-animation* condition as the one with the best performance. The other 3 participants (19%) indicated that the system in the *no-animation* condition performed the best, while the remaining 3 (19%) suggested that both systems had the same performance. None of the participants changed their choice for the *non-reward-based* question. The results are illustrated in Figure 5.1.

5.1.2.2 Reasons for choosing one system over the other – reward-based question

The responses to the question about why participants selected a particular HWR as the one the majority will choose as the best performing one were summarised through thematic analysis. Each response was associated with one or two of the following five themes: *number of errors, generic, type of errors, speed,* and *animation*. Figure 5.2 illustrates the frequencies of these themes. All themes except for *speed* were the same

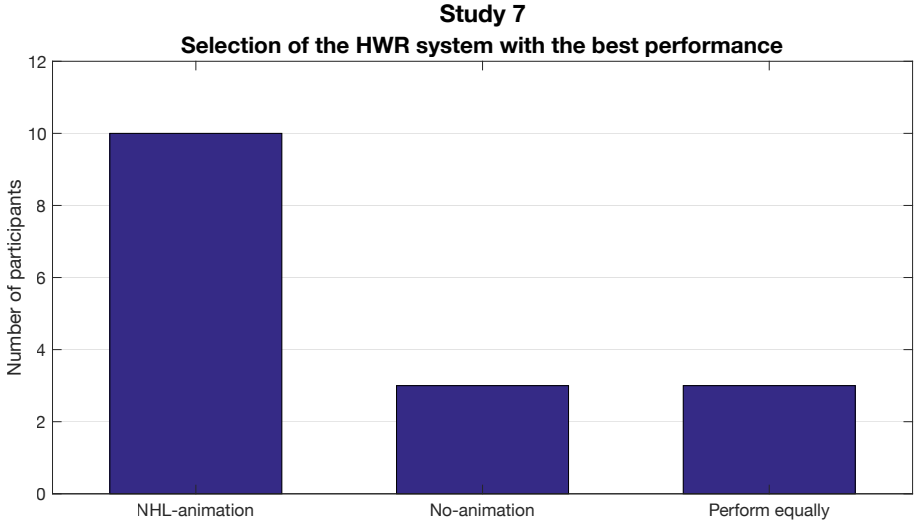

Figure 5.1: Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 7. Number of participants on the y-axis.

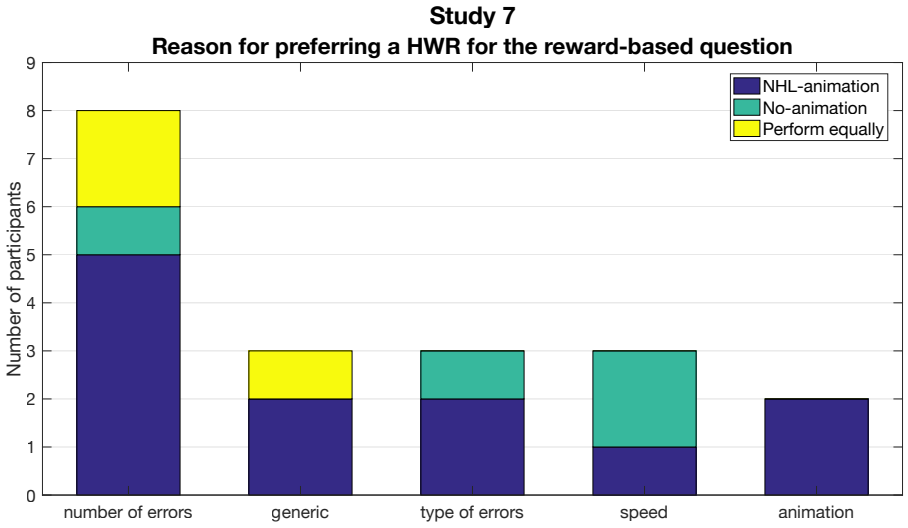

Figure 5.2: Reasons expressed by participants for preferring a HWR in the *NHL-animation* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

as in previous studies. Responses categorised as *speed* are related to comments when participants mentioned that the speed of the system was a reason for their choice. One example of these responses is "Seems that the first program was faster and presented a complete page at once."

5.1.2.3 Reasons for choosing one system over the other – non-reward-based question

Thematic analysis was also applied to the answers related to the non-reward based question. Each response was associated with one or two of the following six themes: *generic, number of errors, speed, type of errors, animation,* and *others' opinion.* Figure 5.3 illustrates the frequencies of these themes. The themes were similar to those which emerged from previous studies. For the majority of participants, 10 out of 16, the answers were the same (in terms of themes) as for the reward-based question. Only 6 participants answered this question differently than the previous one. In particular, the participants who provided different reasons without changing their selection, in one case the answer went from "type of errors and number of errors" to just "generic", in the second case went from "number of errors" to "type of errors". The third case change from "animation" to "generic", in the next case went from "number of errors" to "generic". The fifth case his answer was categorised first as "speed" and changed to "other's opinion". Finally, in the last case went from "type of errors" to just "speed".

5.1.2.4 Performance ratings

A Wilcoxon Signed-rank Test revealed that the performance evaluation was higher in the *NHL-animation* condition ($Mdn = 5$) than in the *no-animation* condition ($Mdn = 4$), ($Z = 2.07, p < 0.05, r = 0.37$). Figure 5.4 shows participant's evaluation of the performance of the systems.

5.1.3 Discussion

The majority of participants in Study 7 selected the system in the *NHL-animation* condition as the one with the best performance, and the Likert-scale ratings for this system were higher, in aggregate than those for the *no-animation* condition, with statistical significance. The analysis of qualitative data is also very much in line with that of previous studies. These results indicate that the effect we observed in previous studies can be observed also for animations that can be interpreted as non-human-like. Therefore the tentative explanation suggested above, that the effect of animations in Studies 4 to 6 may be related to making the system appear more human-like can be rejected. Having rejected human-like explanation, in what follows we turn to the option that the effect of animations may be because the more general relationship between the animation and a user's mental model of the smart system.

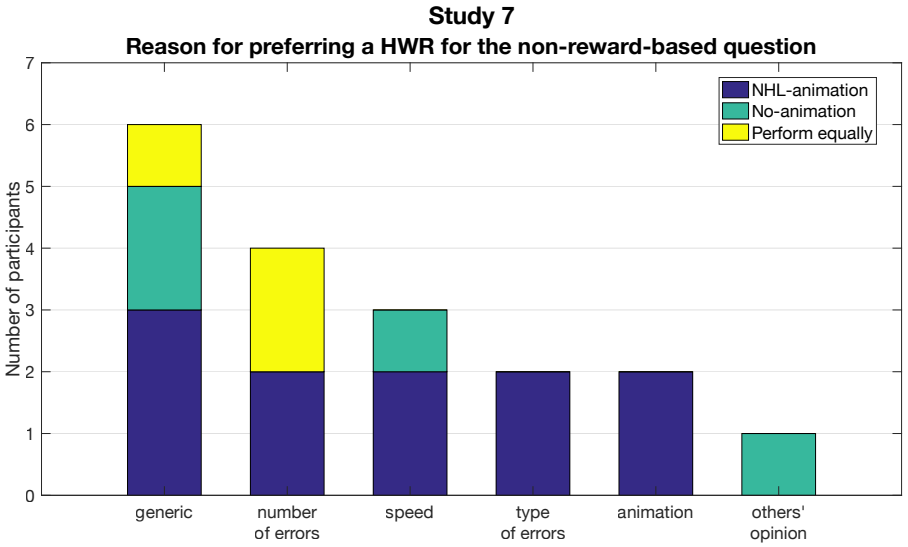

Figure 5.3: Reasons expressed by participants for preferring a HWR in the *NHL-animation* or no-anima*no-animation*tion condition for the non-reward-based question. Number of participants on the y-axis.

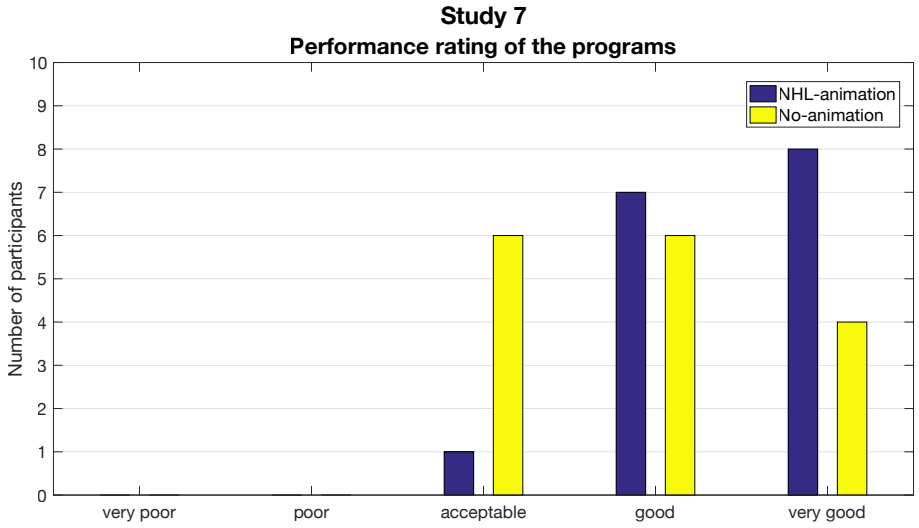

Figure 5.4: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

5.2 Study 8 – People's mental model of HWR system

A new study was designed to investigate the relationship between users' mental models of HWR systems and the animations we displayed in earlier studies. In particular, in this study participants were asked one open question about their idea of how an HWR system works, to check whether these explanations are compatible with the animations used in our prior studies. The study then followed the same structure as Study 5, but at the end, we also asked participants whether their experience of using the system matched their initial idea of how it works.

5.2.1 Method

5.2.1.1 Study Design

The study design was almost identical to Study 5: within-participants, fully counterbalanced, where participants were asked to evaluate and compare the performance of two systems, each corresponding to an experimental condition: *animation* and *no-animation*. The only difference was that we included two additional questions mentioned above.

The reward structure was also identical to Study 5, with a fixed amount being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority. To account for the slightly longer duration of the study (1 extra minute, due to the additional question) the reward for Study 8 was $1.33.

5.2.1.2 Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 5. The sample size was again 16, reported age ranged from 22 to 37 ($M = 29.5, SD = 4.76$), 11 males (69%) and 5 females (31%). All except one were United States nationals, with the remaining 1 South Korean. The education levels of the participants ranged from secondary school level to masters degree level or equivalent. Overall, 2 participants had a master's degree, 9 a university degree, and 5 completed secondary school. None of the participants reported knowing Filipino.

5.2.1.3 Equipment

The same Web application used for Study 5 was used for Study 7, with the only addition of the extra questions, as described above.

5.2.1.4 Procedure

This study followed the same procedure as Study 5, with the exception of the additional open question about how the system works, as described above. The additional question was asked after the initial questionnaire about demographic information and before the main task. In an attempt to prevent random answers, participants were required to submit answers containing at least 20 words. After answering this question participants proceeded to compare the two conditions as in Study 5.

5.2.2 Results

5.2.2.1 How people think an HWR works

The responses to the question regarding how participants think that the HWRs work were analysed through thematic analysis. Two themes emerged in our analysis: *match with database* and *image recognition*. The theme *match with database* included responses that mention using a database to compare the words or characters identified in the handwritten text, such as "The program analyses the written text. It then compares each character to a database loaded into it [...]". The theme *image recognition* was associated with responses that mention how the program processes images to extract characters and words, such as: "It scans the handwriting into an image and then the program look[s] at the image pixel by pixel to match each individual letter [...]'.

Participants' answers suggest that the majority of them seem to have a shared mental model of how the system works. In general, participants agree that somehow the system

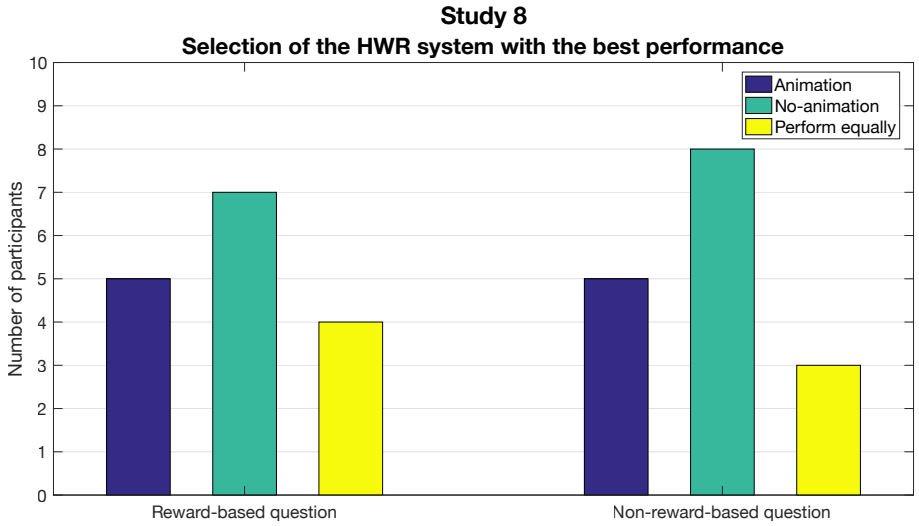

Figure 5.5: Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 5. Number of participants on the y-axis.

has to detect the words or letters to digitise them. Of the 16 participants, 11 stated that the system matches the words and letters using some form of optical recognition. Furthermore, 11 participants mentioned that this then needs to be matched with a 'collection' of some kind, containing labelled examples of handwritten text, finding the corresponding letter or word.

5.2.2.2 Selection of the system with the best performance

For the *reward-based* question, 5 of the 16 participants (31%) chose the system in the *animation* condition as the one with the best performance. Additionally, 7 participants (44%) selected the system in the *no-animation* condition as the one with the best performance, while the remaining four (25%) indicated that both systems had the same performance level. Only 1 participant answered the *non-reward-based* question differently than the *reward-based* one, changing the answer from "both systems" to "no-animation". These results are illustrated in Figure 5.5.

5.2.2.3 Reasons for choosing one system over the other – reward-based question

The responses to the question about why participants selected a particular HWR as the one the majority would choose to have the best performance were summarised through thematic analysis. Each response was associated with one or two of the following five themes: *generic, number of errors, type of errors, animation, speed* and *others' opinion.*

Figure 5.6 illustrates the frequencies of these themes. The themes definitions were the same as in prior studies.

5.2.2.4 Reasons for choosing one system over the other – non-reward-based question

We also analysed why participants considered a particular HWR as the one with the best performance for themselves and concluded with five themes in total: *number of errors, generic, type of errors, animation,* and *speed*. The five themes are the same as above. Figure 5.7 illustrates the frequencies of these themes.

For the majority of participants, 9 out of 16, the answers were the same (in terms of themes) as for the reward-based question. Only 7 participants answered this question differently than the previous one, and of these 7 only 3 changed their selection. In particular, the first participant who selected a different system commented that she believed that other participants would choose the system in the *no-animation* condition because it had the best performance ("I thought the second showed actual better performance"). The second participant considered the majority would select the *animation* condition because this person felt that the other participants would like to see how the system is working. Finally, the last participant mentioned that others would not see a difference between both systems, as such, he selected *both systems* performed equally for the reward-based question and changed to *no-animation* condition for the non-reward-based question. Regarding the other 2 participants who provided different reasons without changing their selection, in one case the answer went from "type of

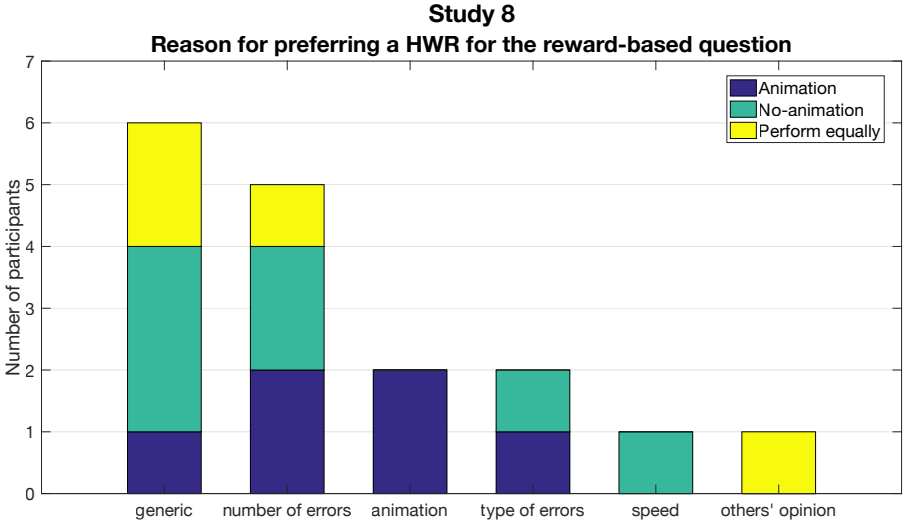

Figure 5.6: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

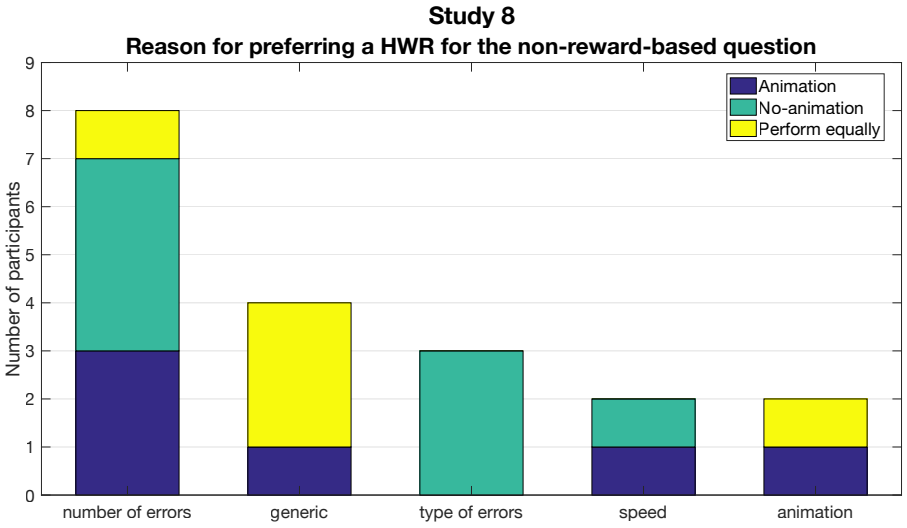

Figure 5.7: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the non-reward-based question. Number of participants on the y-axis.

errors and number of errors" to just "number of errors", on the other case, it went from "number of errors" to "animation" ("I liked that the second computer program highlighted the text in red as it was transcribing it").

5.2.2.5 Performance ratings

A Wilcoxon Signed-rank Test revealed no significant difference in the evaluation of the performance between the *animation* condition and *no-animation* condition ($Z = 1.03, p = 0.31, r = 0.18$). Figure 5.8 shows participant's evaluation of the performance of the systems.

5.2.2.6 The system worked as participants expected.

All participants reported that both systems successfully transcribed the handwritten text to typed text, and so they considered that the systems worked as they expected. Additionally, only three participants mentioned in their comments the animation (e.g. "For the second program, it showed how the program scanned each word in red. It was computing for the e-text").

5.2.3 Discussion

The explanations that participants provided about how an HWR system works matched quite closely how this kind of systems are actually implemented: they perform some form

of image recognition on characters and words. All participants seem to have an accurate mental model regardless of whether they have formal technical background (indeed only 4 of them did). Moreover, the explanations participants provided seem to be quite in line with the animation that we implemented for Studies 4, 5 and (to an extent) 7. The match, however, is not always an exact one: 14 out of the 16 participants explained that the recognition would happen letter by letter, so in the same way a human would actually type handwritten text into a computer. In contrast, the animation implemented in the previous studies can be interpreted as processing the text word by word rather than character by character.

The results from Study 8 seem to be in stark contrast to those from Studies 4 to 7. Only 5 participants selected the system in the *animation* condition as the one with the best performance level, and the analysis of the Likert-scale ratings did not reveal statistically significant differences between the conditions, despite the sample size being the same as in the earlier studies. The different results can be attributed to the additional question about participants' mental model of HWR systems asked at the beginning of the study.

Arguably, asking participants how they think an HWR system works makes their mental model for this kind of system salient to them. This salience seems to contrast the effect of the animation that we observed in earlier studies. Perhaps, then, making participants aware of how the system works has an effect similar to that of the animation in our earlier studies. In other words, these results suggest that in our earlier studies the animations reminded participants of how the smart system works, instead of in Study 8 the preliminary question had the same effect, so it seems to have replaced the effect of the animation (for both conditions). This finding resonates with studies in Psychology

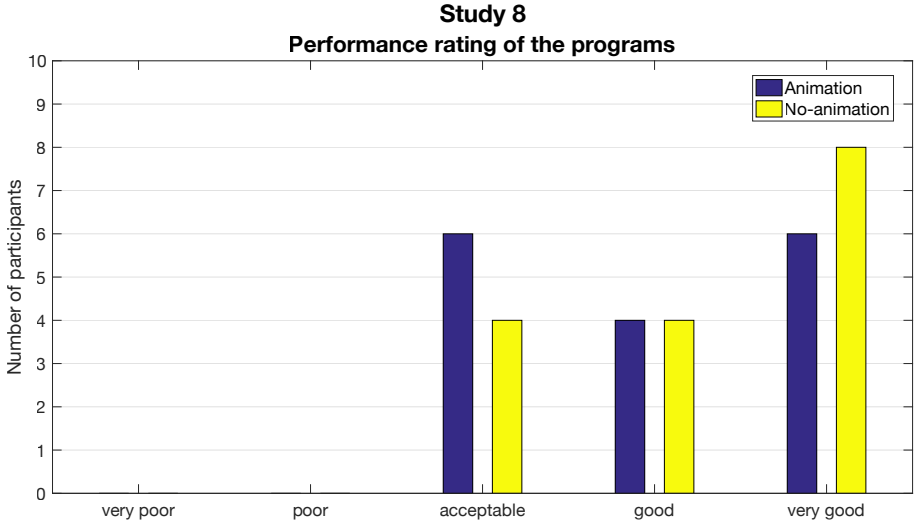

Figure 5.8: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

which demonstrated that making a bias salient to participants may remove the effect of the bias. In particular Schwarz and Clore (1983) through a well-known study about the effect of weather on mood demonstrated that asking participants about the weather (and hence making the weather salient to them) removes the effect that weather has on mood (at least in the case of bad weather). Similarly, in our study asking participants about how the system works seems to remove the effect of the animation. We further explore the relationship between mental models and the effect of animations on perceptions of performance in the following study.

5.3 Study 9 – Mental model match and mismatch animations cues vs. no-animation

The results of Study 8 suggest that the animation used in Studies 4 and 5, and to some extent in Study 7, matched participants' mental models of HWR systems. Based on such finding, a possible explanation of the results from Studies 4, 5 and 7 is that the animations we displayed "reassured" participants that they system work as they expected. As such, the animation raised their confidence in the system and enhanced their perception of its performance.

Based on this, we formulated an explanation of how an HWR system works which matches the animation that we used in Studies 4 and 5. We refer to this one as the *original* animation. We also designed a new animation, which we refer to as the *alternative* animation, and we formulated a corresponding explanation. The *alternative* animation consisted of enclosing each word with a rectangle and inverting its colour, before displaying the corresponding word on the right-hand side of the screen[2]. This alternative animation was designed to be at odds with the explanations collected from participants in Study 8 about how an HWR works. For consistency, both animations, original and alternative, included the same transcription errors. Study 9 was designed to compare and contrast the effect of *matching* and *mismatching* animations and explanations.

5.3.1 Method

5.3.1.1 Study Design

The two animations and the two explanations described above define 4 conditions in a 2×2 fashion: (original animation, original explanation), (original animation, alternative explanation), (alternative animation, alternative explanation), and (alternative animation, original explanation). In the first and third condition, animation and explanation are *matching*, while in the second and fourth they are *mismatching*. These 4

[2] https://vimeo.com/183480642.

conditions were applied through a between-participants study design: each participant was assigned to one of these 4 conditions. Similar to Study 5, each participant was asked to evaluate and compare the performance of two HWR systems: one involving an animation (*animation* condition) and one with no animation (*no-animation*). The *no-animation* condition, which was similar to the condition in Study 5, was the same for all participants. The *animation* condition would involve either the *original* or the *alternative* animation, depending on the assigned group of the participant. The *animation* and *no-animation* conditions were presented to participants in fully counterbalanced order.

The reward structure was identical to Study 5, with a fixed amount being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority. To account for the slightly longer duration of the study (compared to Study 5) the reward for Study 9 was $1.33.

5.3.1.2 Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 5. The sample size was 64, higher compared to earlier studies, to reflect the larger number of conditions and the between-participants design. Age ranged from 20 to 61 ($M = 32, SD = 11.22$), 34 males (53%) and 30 females (47%). All except for 2 of the participants reported to be United States nationals, the remaining two being Canadian and South Korean. The education levels of the participants ranged from secondary school level to doctoral degree: 1 participant had a doctoral degree, 5 had a masters' degree, 39 a university degree and 19 completed secondary school. One of the participants reported knowing Filipino.

5.3.1.3 Equipment

The Web application used for Study 5 was modified to include the two different animations described above, and to include an initial explanation of how the system works, together with a *reinforcing question*, as detailed below. As in Study 5, the Web application was deployed to a publicly accessible Web server, to allow MTurk' workers to access it from their personal computers. Two additional questions were also included in the final questionnaire, to ask participants whether they considered that the systems worked according to the explanation they received at the beginning of the study and why.

5.3.1.4 Procedure

In addition to the procedure followed in Study 5, the explanation of how the system works (*original* or *alternative*, depending on the condition) was presented to the participants.

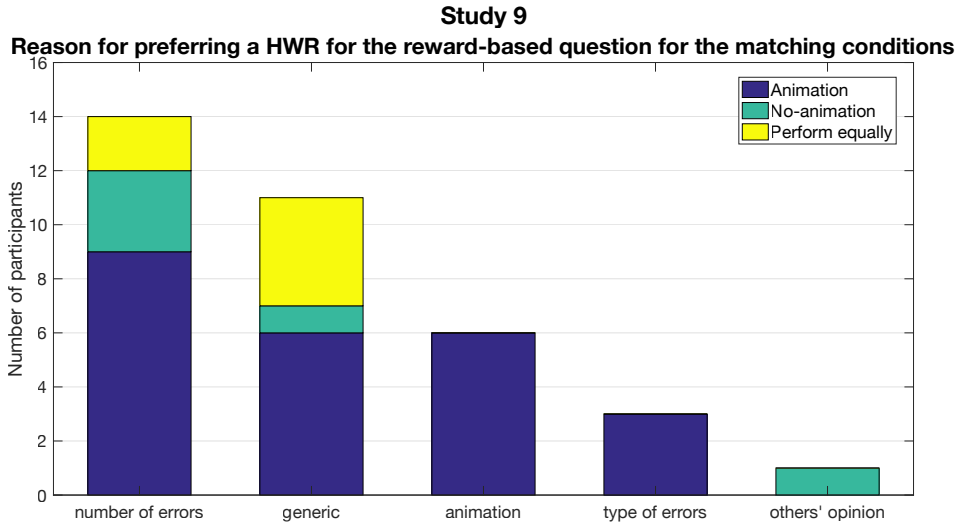

Figure 5.9: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question for the match conditions. Number of participants on the y-axis.

The explanation was given to the participants after the questionnaire where we asked them about their personal information, with the aim of influencing their mental model of the system. However, participants were not explicitly told that the explanation would correspond to any *animation cues*. After the initial explanation, participants were asked to explain, in their own words, how they think the systems works – we refer to this as the *reinforcing question*. Participants were told that their responses should have more than 20 words to be considered valid. Subsequently, the two HWR systems were presented to the participants. Similar to the previous studies, participants were then asked to evaluate the performance of the systems on a 5-point Likert scale, to select the system they considered to perform best (or that the systems had the same performance level) and to justify their selection. At the end of the study, after evaluating both systems, participants answered the two new questions that we designed.

5.3.2 Results

5.3.2.1 Selection of the system with the best performance

Matching conditions. For the *reward-based* question, 21 of the 32 participants (66%) selected the system in the *animation* condition as the one with the best performance. Additionally, 5 participants (15%) selected the system in the *no-animation* condition as the one with the best performance, while the remaining 6 (19%) indicated that both systems had the same performance level. None of the participants answered the *non-reward-based* question differently than the *reward-based* one. The participant who reported knowing

Filipino was in this condition, and she selected the system in the *animation* condition as the one with the best performance for both questions.

Mismatching conditions. For the *reward-based* question, 10 of the 32 participants (31%) selected the system in the *animation* condition as the one with the best performance. Additionally, 12 participants (38%) selected the system in the *no-animation* condition as the one with the best performance, while the remaining 10 (31%) indicated that both systems had the same performance level. 16 of the participants answered the *non-reward-based* question differently than the *reward-based* one.

5.3.2.2 Reasons for choosing one system over the other – reward-based question

Participants' responses to the question about why they chose a particular HWR system as the one that the majority will choose to have the best performance were categorised through thematic analysis. Each response was associated with one or two themes, with five themes found in total: *number of errors, generic, animation, type of errors, speed* and *others' opinion*. The five themes are the same as in Study 5. Figure 5.9 illustrates the frequencies of these themes, also classified on the reward-based question (*animation* or *no-animation*) for the **matching conditions**, while Figure 5.10 illustrates the frequencies for the **mismatching conditions**. The participant who reported knowing Filipino was in the matching condition, and her answers were associated with the theme *animation* for both questions.

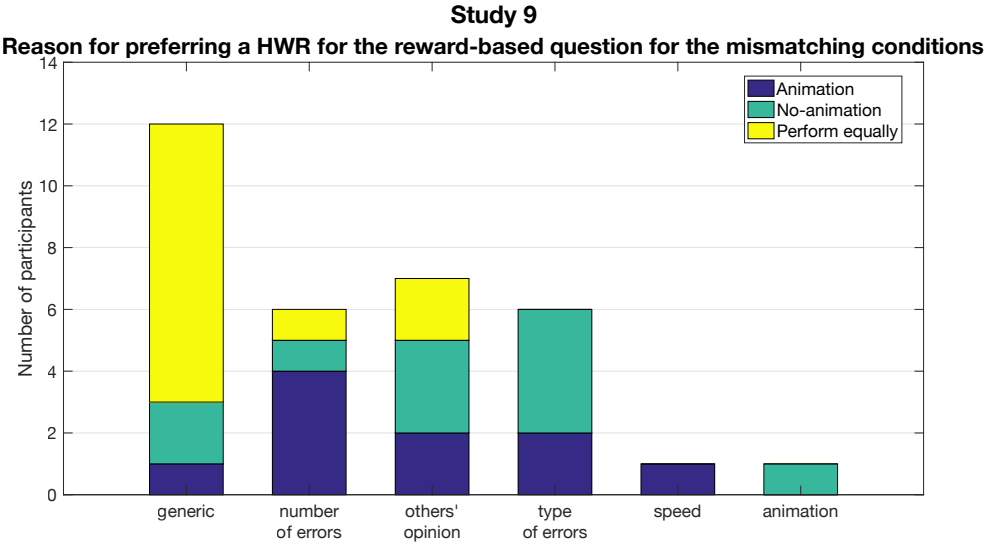

Figure 5.10: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question in the match conditions. Number of participants on the y-axis.

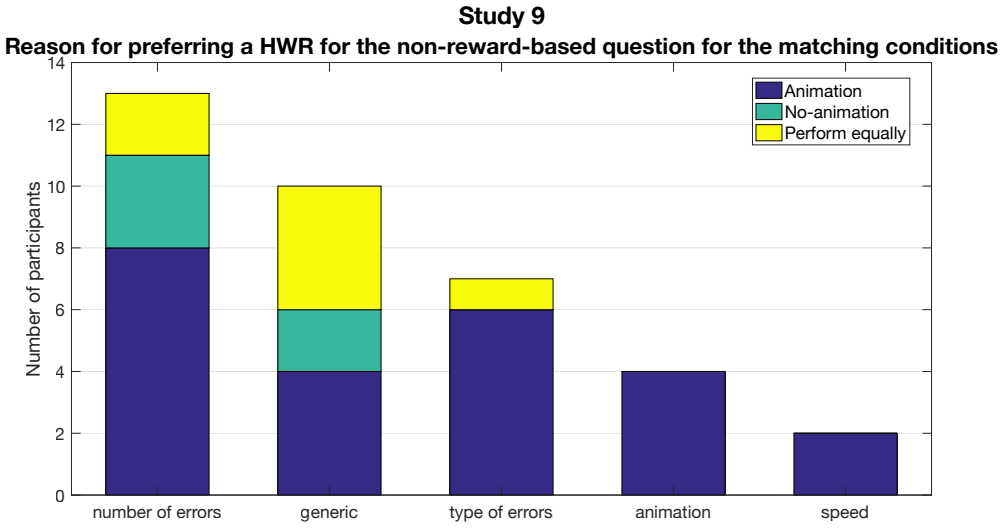

Figure 5.11: Reasons expressed by participants for preferring a HWR in the animation or *no-animation* condition for the non-reward-based question in the match conditions. Number of participants on the y-axis.

5.3.2.3 Reasons for choosing one system over the other – non-reward-based question.

Participants' responses to the question why they selected a particular HWR as the one with the best performance were categorised through thematic analysis. Each response was associated with one or two themes, with five themes found in total: *number of errors, generic, type of errors, animation* and *speed*. The themes categorised participants' comments as we did in the previous studies. Figure 5.11 illustrates the frequencies of these themes, also classified on the individual preference (*animation* or *no-animation*) for the **matching conditions**, while Figure 5.12 illustrates the frequencies for the **mismatching conditions**.

5.3.2.4 System working according to expectations

Matching conditions. Overall, 27 of the 32 participants indicated that the systems worked as they expected from the explanation given at the beginning of the study, while the remaining 5 stated that it did not. In more detail, participants considered that the system compares each word with a database. As such, 3 participants believed the system would process the data word by word rather than character by character as the errors they found (e.g. "It seemed the program did it letter by letter, not by the word as described. But then again I don't know the language, so changing one letter like the programs did may still have been recognizing a word."). Other 2 participants mentioned that they believe that the systems did not work because only one system shows the animation (e.g. "The second computer program highlights the words as it

transcribes them. The first didn't appear to do that."). In the free text comments, 10 participants mentioned the animation explicitly as a reason of why they considered the system worked as they expected (e.g. "I could see the text being highlighted and picked apart", and "Because you could see the process of transcription as it was happening.").

Mismatching conditions. Overall, 25 of the 32 participants indicated that the system worked as expected from the explanation given at the beginning of the study, while the remaining 7 stated it did not. In the free text comments, 10 participants mentioned the animation. In more detail, 3 participants mentioned that the animations mismatched the explanation. Moreover, the 3 participants who reported that the system did not work as they expected mentioned that they thought the system transcribed the handwritten text word by word rather than character by character because of errors they found in the typed text. In contrast, other participants felt that the system transcribed the handwritten text better than they expected. Because of this, they felt that the system worked correctly.

Participants' responses to why they considered the that the systems worked (or not) according to their expectations were categorised through thematic analysis. Each response was associated with one or two themes, with eight themes in total: *generic, animation, faith, disbelief, correctness, analysis, experience,* and *technology.* Figure 5.13 illustrates the frequencies of these themes for the **matching conditions**, while Figure 5.14 illustrates the frequencies for the **mismatching conditions**. The theme generic was associated with responses where participants did not provide full explanation or misunderstood the question, such as "I believe they translated the handwritten text into a digital computer font". The theme *animation* was used when participants talked about

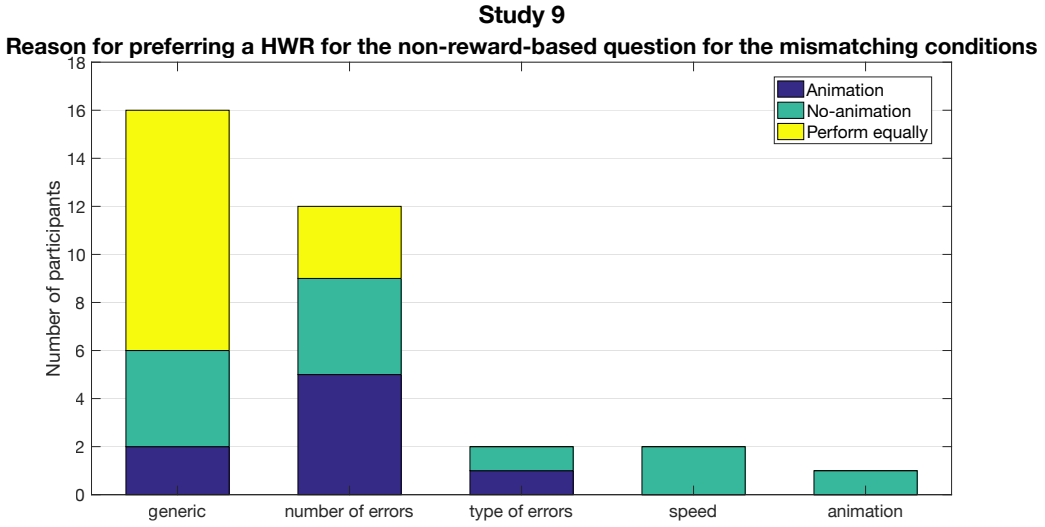

Figure 5.12: Reasons expressed by participants for preferring a HWR in the animation or *no-animation* condition for the non-reward-based question for the mismatch conditions. Number of participants on the y-axis.

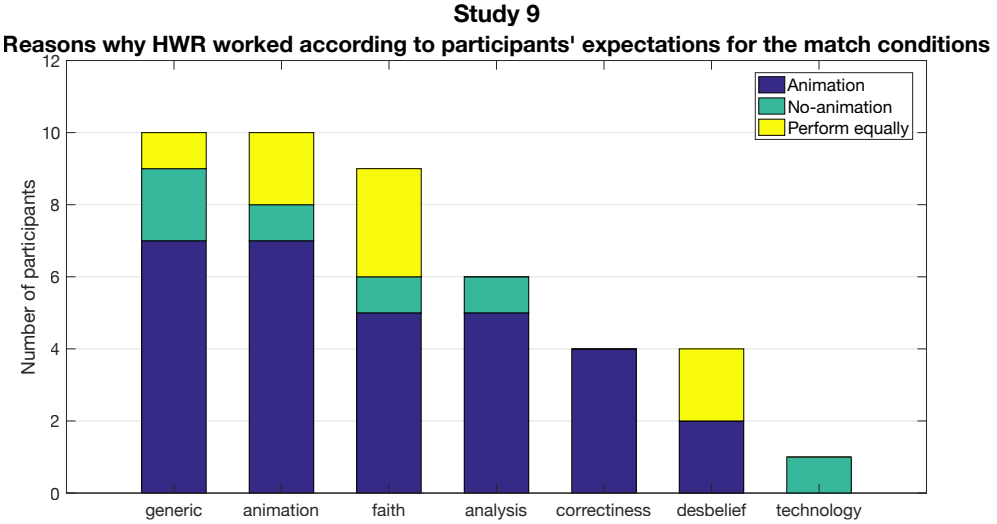

Figure 5.13: Reasons expressed by participants for why they considered that the systems worked according to their expectations for the match conditions. Number of participants on the y-axis.

why the animation affected their consideration of whether the systems are working or not, such as: "Because you could see the process of transcription as it was happening". We grouped a response into the theme *faith* if it is related to the participants believing the explanation provided: "I had no reason to doubt the explanation, it seemed perfectly reasonable". Comments grouped into *disbelief* is the opposite, and instead it's related to situations where the participants do not believe in the explanation provided: "I don't see how changing the color of the handwritten text to match the color of the paper as a way to convert the text to etext [...]". Responses related to the accuracy of the output, such as "Yes, it translated the characters of the handwritten text correctly", were categorised as *correctness*. The theme *analysis* was used to categorise comments that talk about the actual transcription process, such as "It appears that the program goes through each letter and tries to identify which letter it is". The theme *experience* was used for any comments in which the participant talk about his/her own personal experience with HWR systems: "I've used OCR [Optical Character Recognition] programs before and they were never as accurate as this one was, so I don't believe it actually exists". Finally comments such as "Technology and artificial intelligence is growing at an exponential rate" were categorised as *technology*.

5.3.2.5 Performance ratings

Matching conditions. A Wilcoxon Signed-ranks Test revealed that the performance evaluation was higher for *animation* condition ($Mdn = 5$) than for *no-animation* condition ($Mdn = 4$), ($Z = 2.94, p < 0.01, r = 0.37$). Figure 5.15 shows participant's evaluation of the performance of the systems.

Mismatching conditions. A Wilcoxon Signed-ranks Test did not reveal statistically significant differences. Figure 5.16 shows participant's evaluation of the performance of the systems.

5.3.3 Discussion

The results of this study confirm what was suggested by the findings of Study 8: *animation cues* influence the perception of the system performance only if they are compatible with the participant's mental model. More in general, taken together with the results of Study 8, these results allow us to propose the following explanation for the effect:

We used the word 'largely' because of the results of Study 8, indicate that the explanations for how the system works provided by participants do not match *exactly* the animation. However, if the explanation is radically different, as in the *mismatching* conditions of Study 9, the effect disappears.

Moreover, as mentioned above, the results of Studies 4 to 7 suggest that this effect takes place largely unconsciously – most participants did not mention the animation in their justification for the selection of the system with the best performance. Similarly, in Study 8 we observed that if the explanation of how the system works is made salient to participants, in that case through an initial question, the effect of the animation disappears. Arguably, the explanations that participants provided in Study 8 could apply to *both* the systems they evaluated, so that process reminded them, or made them aware, of how *both* systems work. Hence, their judgement was not biased towards either of them. It should be noted that in Study 9 participants were also exposed to

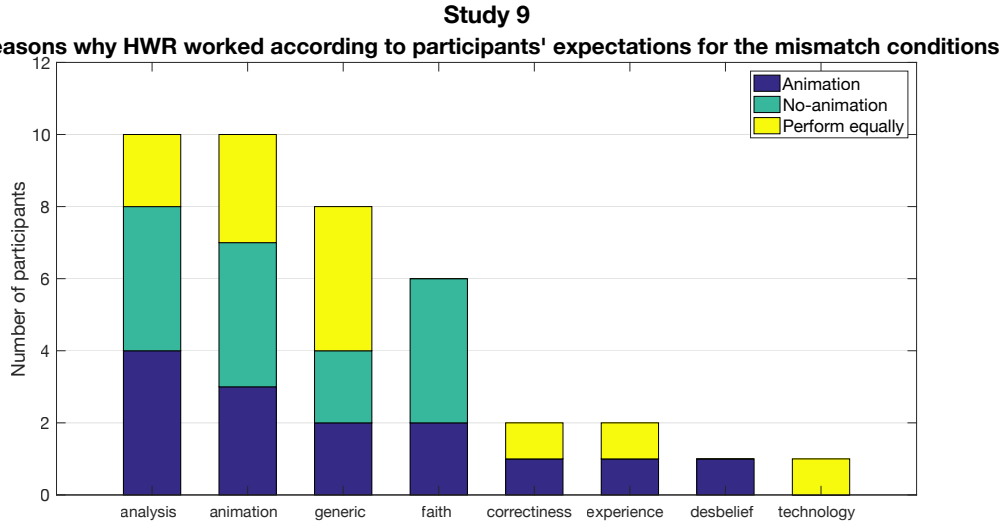

Figure 5.14: Reasons expressed by participants for why they considered that the systems worked according to their expectations for the mismatch conditions. Number of participants on the y-axis.

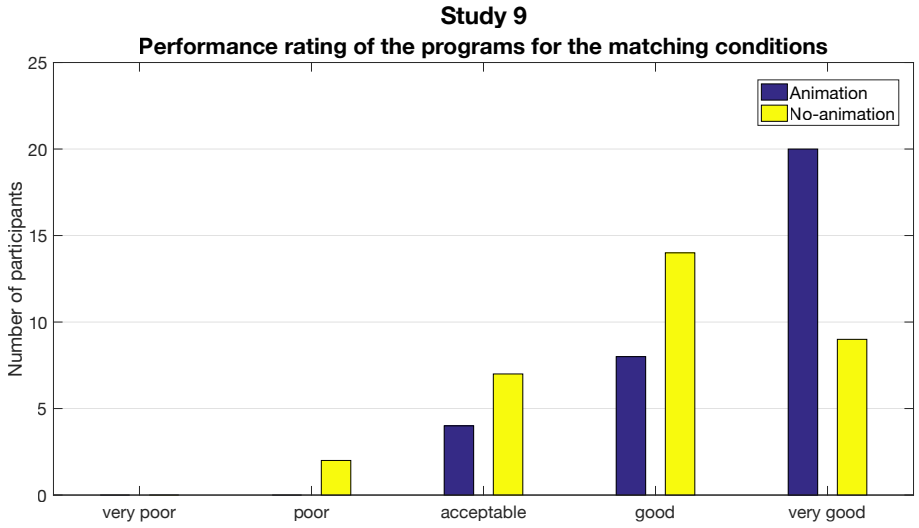

Figure 5.15: Likert-scale of participants' evaluation of the performance of the systems in the *animation* and *no-animation* conditions in the matching conditions. Number of participants on the y-axis.

an explanation of how an HWR system works at the beginning of the study. However, in that case, the explanation was provided to them, and it matches very closely the animation shown – these differences may explain why the effect of animation was still observed in this study compared to Study 8.

In fact, 13 participants (out of 64) mentioned that the reason they think that the system works (or not) is largely based on the explanations they received.

Even though the importance of mental models in HCI has been discussed for at least three decades (Kieras and Bovair (1984); Norman (2013)), most prior work focussed on the effect of mental models on *users' performance* when using an interactive system. In contrast, our results suggest a relationship between mental models and users' perception of the *system's performance*.

In the qualitative data, we did not find comments of people explaining that they perceived a match or mismatch between the explanation that elicits a mental model and the animation they received. This behaviour suggests that the effect of *animation cues* happens unconsciously. Thus, we can argue that the participants' comments and evaluation make visible that indeed the fact that the animation matches participants' mental model affects their perception. The participant who reported knowing Filipino selected the same answers as the majority of other participants (the system in the *animation* condition as the one working better), so her knowledge of the language does not seem to play a visible role here. She justified her selection referring to the animation – perhaps her language knowledge allowed her to give more attention to this feature of the UI, compared to other participants who might have busy comparing the input and output portions of the interface. However, further work would be required to evaluate the effect

of language familiarity on the effect of animation. Now that we found the reason behind why the *animation cues* influence people's perception on how they perceive smart systems' performance, we move to further characterise this effect with the following studies.

5.4 Summary

In this section, we focused on understanding which characteristics the *animation cues* require assuring the effect on people's perception persist when designers implement *animation cues* as a *visual feedback* in their systems. In Study 7 (N=16), we found that the findings indicate the effect we observed in Study 4, 5, and 6 can also be observed for animations that can be interpreted as *non-human-like*. Thus, we ran Study 8 (N=16) to investigate the relationship between users' mental models of HWR systems and the animations we displayed in the studies of Chapter 4. We observed that the explanations that participants provided about how an HWR system works matched quite closely how this kind of systems are implemented: they perform some form of image recognition on characters and words. Moreover, the explanations participants provided seem to be quite in line with the animation that we implemented for Studies 4, 5, and 7.

Based on Study 8 findings, we designed Study 9 (N=64) to analyse whether the effect we found in previous studies was because the animation we designed to match people's mental model. The findings of this study confirm that *animation cues* influence the perception of the system performance only if they are compatible with people's mental model. Thus, we suggest that designers need to consider that the design of *animation*

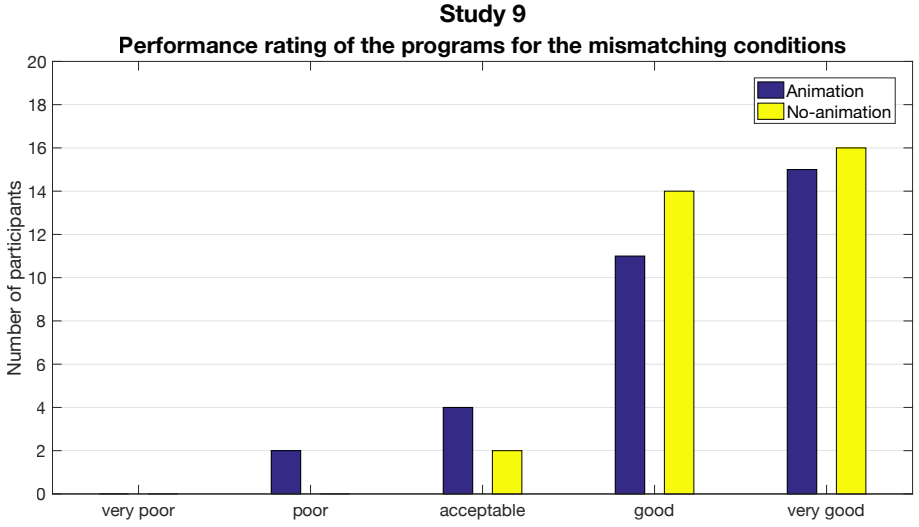

Figure 5.16: Likert-scale of participants' evaluation of the performance of the systems in the *animation* and *no-animation* conditions in the mismatching conditions. Number of participants on the y-axis.

cues as a *visual feedback* have to match people's mental model of how their systems perform their task.

Chapter 6

The Effect of Varying Other Dimensions of the Animation Cues

From the studies of Chapter 4 and 5, we learned that *animation cues* have a positive effect on people's perception of smart system's performance. In addition to these findings, we wanted to analyse whether the varying of other dimensions (e.g. amount of detail in the animation, the number of errors the system has or animation's speed) of the *animation cues* can affect their effect on people's perception. Hence, in what follows, we detail three studies to investigate this. In more detail, Study 10 (N=16) was designed to address the research question "Can higher detail of animations better influence people's perception of smart systems than a lower detail of animations?". In addition, in Study 11 (N=32) we tested our research question "What level of imbalance in the performance level of the system being compared would "break the illusion" created by the *animation cues*?". Finally, we designed Study 12 (N=48) to address the following research question: "Does varying the speed of the animations would "break the illusion" created by the *animation cues*?"

6.1 Study 10 – Detail of animation

Through Study 4 and 5, we found that *animation cues* can influence how people evaluate the performance of screen-based systems. As a subsequent step, we evaluate whether the amount of detail of a displayed motion, so how much animation needs to be shown in all the elements related to the system's task (e.g. handwritten text and e-text), can have an impact on participants' perception. We expect to find a relationship between the amount of motion displayed and the perceived performance.

To explore this, we designed a new animation, which involves less motion than the animations used in previous studies.

6.1.1 Method

6.1.1.1 Study Design

The study design was almost identical to Study 5: within-participants, fully counterbalanced. However, it involved 3 conditions: *animation*, *partial-animation*, and *no-animation*. Similar to prior studies, each condition corresponded to a system that participants were asked to evaluate and compare in terms of performance. The new *partial-animation* condition is similar to the *animation* condition, except that instead of involving the animation on both the input and output parts of the UI (i.e. on both the handwritten and typed text), it only applies to the output part of the UI (approximately the right half of the screen), while the input part of the UI remains static.

The reward structure was also identical to Study 5, with a fixed amount being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority. To account for the slightly longer duration of the study (compared to Study 5) the reward for Study 10 was $2.

6.1.1.2 Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 5. The sample size was 48, to account for the increased number of conditions, reported age ranged from 22 to 44 ($M = 34, SD = 9.53$), 31 males (65%), 15 females (31%), 1 prefer not to say (2%), and 1 other (2%). Of those participants, 46 were United States nationals, 1 was Polish, and the other was British. The education levels of the participants ranged from primary school level to masters' degree level or equivalent. Overall, 2 participants had a master's degree, 30 a university degree, 15 completed secondary school, and 1 completed primary school. Two of the participants reported knowing Filipino.

6.1.1.3 Equipment

The Web application used for Study 5 was modified to include the additional condition described above.

6.1.1.4 Procedure

This study followed the same procedure as Study 5, with the exception of the additional condition outlined above.

6.1.2 Results

6.1.2.1 Selection of the system with the best performance

For the *reward-based* question, 26 of the 48 participants (54%) selected the system in the *animation* condition as the one with the best performance, 9 participants (19%) selected instead the system in the *partial-animation* condition, 7 participants (15%) the system in the *no-motion* condition, while the remaining 6 participants (12%) suggested that the three systems had the same performance. Moreover, 5 participants changed their choices for the *non-reward-based* question as follows: from 'Partial-animation' to 'Animation', from 'All performed equally' to 'Animation', from 'Animation' to 'No-animation', from 'All performed equally' to 'No-animation', and from 'Animation' to 'All performed equally'. These results are illustrated in Figure 6.1.

6.1.2.2 Reasons for choosing one system over the others – reward-based question

We categorised participants' responses about why they selected a particular HWR as the one with the best performance for the reward-based question. Each response was

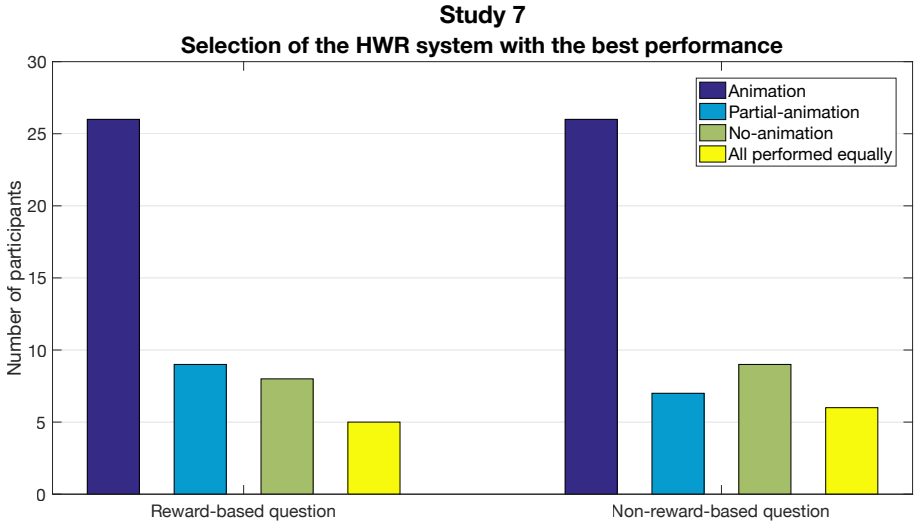

Figure 6.1: Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 10. Number of participants on the y-axis.

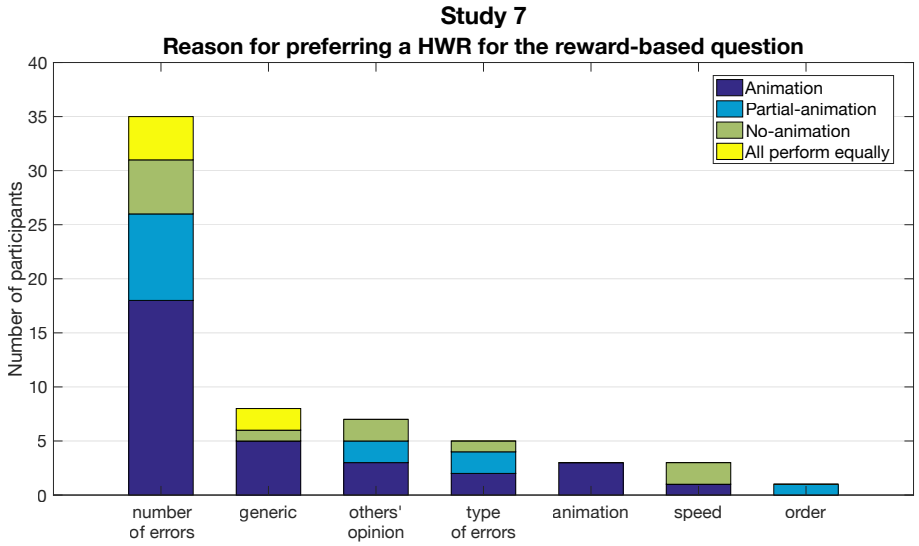

Figure 6.2: Reasons expressed by participants for preferring a HWR in the *animation*, *partial-animation* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

associated with one or two of the following seven themes: *number of errors, generic, others' opinion, type of errors, animation, speed,* and *order*. Figure 6.2 illustrates the frequencies of these themes. The themes were the same as those emerged from previous studies, with the exception of *order* which was used for the following comment: "I didn't find any errors in the first program, and it is the first on the list."

6.1.2.3 Reasons for choosing one HWR system over the others – non-reward-based question

The comments of the participants about why they selected a particular HWR as the one with the best performance for the non-reward-based question were categorised through thematic analysis. Each response was associated with one or two themes, with seven themes used in total: *number of errors, type of errors, generic, speed, animation, others' opinion,* and *random*. Figure 6.3 illustrates the frequencies of these themes. While the first six themes are the same as above. The new theme *random* was associated with the response "I am not sure, all of them seemed to perform similarly, but it looks like the third was maybe the best? Not sure."

For the majority of participants, 25 out of 48, their answers were the same (in terms of themes) for the reward-based question. Only 23 participants answered this question differently than the previous one, and of these 23 only 4 changed their selection. In more detail, the first participant who selected a different system commented that he believed that other participants would choose the system in the *partial-animation* condition because it was the first system they saw and they would remember ("I think

people might choose the first one because their attention and focus will be greater when viewing the first program compared to the second and third."). The second participant considered the majority would select that the *three systems* have the same performance for the reason that had the same performance. However, he felt that the system in the *no-animation* condition was the one with the best performance because it was the only one that produced all of its output in one go. The third participant considered that other participants would be distracted to evaluate the system in the *animation* condition ("People will lose some concentration, so they will not be accurate.") Finally, the last participant changed their selection from *three systems* performed equally to the system in the *animation* condition. Regarding 19 participants who provided different reasons without changing their selection, 4 participants changed their answers from "number of errors" to "generic". Additionally, 3 participants' answers changed from "number of errors" to "type of errors", 2 participants' answer changed from "generic" to "generic and type of errors", 2 participants' answers change from "generic" to "speed". Moreover, three participants' answers changed from "others' opinion" to "random" to "number of errors" to "type of errors". In another case, two participants changed their answers from "type of errors" to "number of errors" to "speed". In another case the answer was changed from "speed" to "generic", and on another, the answer changed from "generic" to "number of errors". Finally, in the last case, the participant changed the answer from "number of errors and generic" to just "number of errors".

For this study, five participants changed their choice of which HWR system performed the best. The two participants that changed their choice from *partial-animation* and three programs to *animation* condition mentioned that they considered that the *animation* condition has the best performance and so other participants will pay more

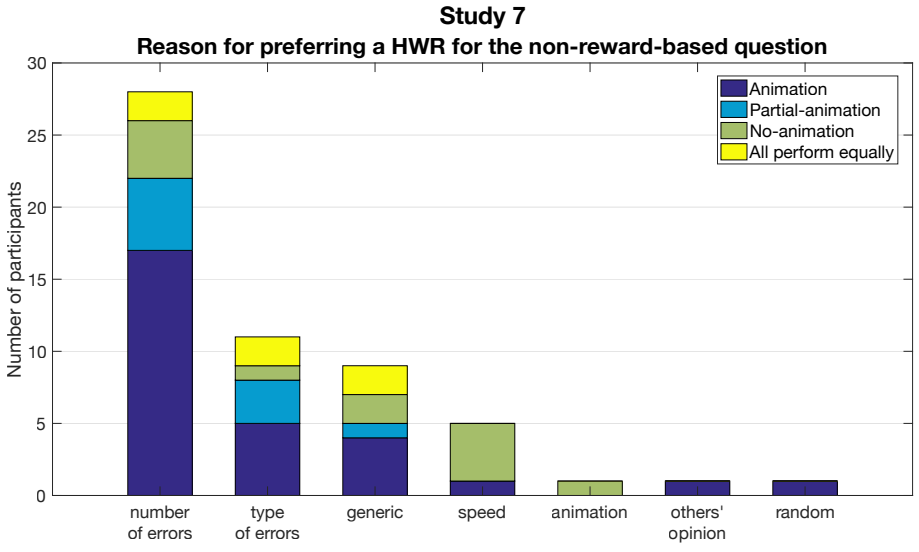

Figure 6.3: Reasons expressed by participants for preferring a HWR in the *animation*, *partial-animation*, or *no-animation* condition for the non-reward-based question. Number of participants on the y-axis.

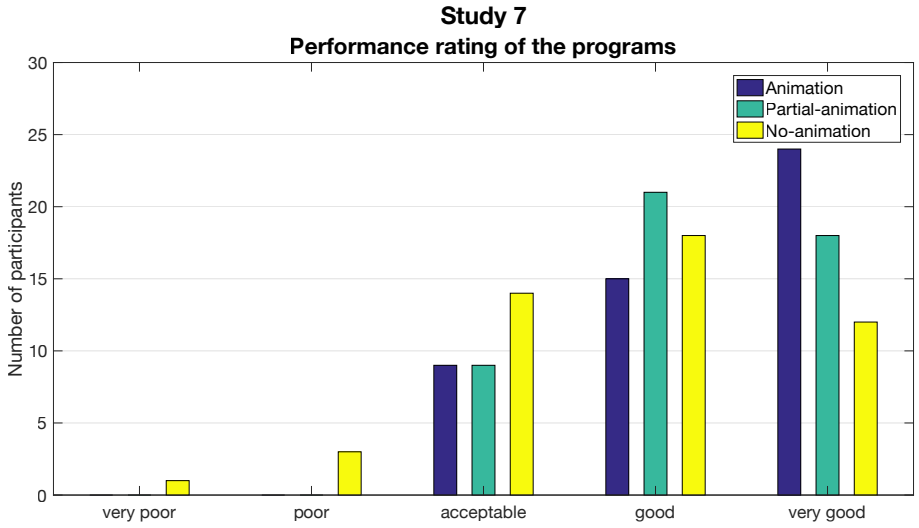

Figure 6.4: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

attention to the system in that condition. The two participants that changed their choice from *animation* and all performed equally to *no-animation* commented that they considered that this condition is faster than the other condition and this will have motivated people to select the other conditions. Finally, the participant that changed from *animation* condition to all performed equally considered that most of the participants will select the animation because this will distract them.

6.1.2.4 Performance ratings

A CHI-squared test revealed statistically significant differences in the performance ratings, $\chi^2(2) = 9.73, p < 0.01$. Post-hoc analysis with pairwise Wilcoxon Signed-ranks tests with significance level set at $p < 0.05$, revealed Median for the performance evaluation for the *animation*, *partial-animation*, and *no-animation* were $Mdn = 4.5$, $Mdn = 4$, and $Mdn = 4$, respectively. There were significant differences between the *animation* and the *no-animation* conditions ($Z = 3.037, p < 0.01, r = 0.31$), and also between the *partial-animation* and the *no-animation* conditions ($Z = 2.64, p < 0.05, r = 0.25$). No significant differences were instead found between the *animation* and *partial-animation* conditions ($Z = 0.89, p > 0.05, r = 0.09$). Figure 6.4 shows participant's evaluation of the performance of the systems.

6.1.2.5 Discussion

The results of Study 10 suggest that *any amount* of animation seems to influence users' perception of the performance of the system: statistically significant differences in the

Likert-scale ratings were found both between *no-animation* and *animation* and between *no-animation* and *partial-animation*, while no statistically significant differences were found between *animation* and *partial-animation*. However, in terms of choosing the system with the best performance, the majority of participants opted for the *animation* condition, rather than any of the other 3 options, regardless of the experimental reward. Indeed, almost three times as many participants opted for the system in the *animation* condition compared to the one in the *partial-animation* one. In contrast to the Likert-scale results, the selection results suggest that the amount of animation does play some role in users' perception of performance. So perhaps the lack of statistical significance mentioned above could be a limitation of our sample size.

6.2 Study 11 – Decreasing performance

In all studies reported so far, the presence of animation was the only difference between the systems our participants evaluated. The performance of the various systems, defined in terms of number of errors produced by the system was kept constant. To further characterise the effect we identified, we decided to test what level of imbalance in the performance level of the system being compared would "break the illusion" created by the animation. In other words: how many additional errors can the animation cover? Study 8 was designed to address this question, by comparing pairs of systems with different numbers of mistakes.

6.2.1 Method

6.2.1.1 Study Design

The study design was based on the design of Study 5, but we additionally divided participants into two groups, each group corresponding to a different number of errors: *9-errors* group and *10-errors* group. For each group, the experiment was identical to Study 5, with the exception that the *animation* condition the number of errors indicated by the group name. The *no-animation* condition always included just 8 errors, as it was in Study 5 (in earlier studies the number of errors was the same across the conditions).

The reward structure and amounts were identical to Study 5, with a fixed amount of $1.17 being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority.

6.2.1.2 Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 5. The sample size was 32, double than what it was for Study 2, to account for the split of participants into 2 groups. The age ranged from 19 to 62 ($M = 27, SD = 6.80$), 22 males (69%) and 10 females (31%). All except for 3 of the participants reported to be United States nationals, the remaining ones being Algerian, Canadian and Japanese. The education levels of the participants ranged from secondary school level to masters' degree level or equivalent. Overall, 6 participants had a masters' degree, 20 had a university degree, and 6 completed secondary school. One of the participants reported knowing Filipino.

6.2.1.3 Equipment

The same Web application used for Studies 4 and 5 was used for Study 11, with the only difference of the number of errors in the *animation* condition, as described above.

6.2.1.4 Procedure

The procedure was the same as Study 5.

6.2.2 Results

6.2.2.1 Selection of the system with the best performance

9-errors group. For the *reward-based* question, 9 of the 16 participants (56%) chose the system in the *animation* condition as the one with the best performance. Additionally, 4 participants (25%) selected the system in the *no-animation* condition as the one with the best performance, while the remaining 3 (19%) indicated that both systems had the same performance level. Only 1 participant answered the *non-reward-based* question differently than the *reward-based* one, changing the answer from "animation" to "both systems". These results are illustrated in Figure 6.5.

10-errors group. For the *reward-based* question, 4 of the 16 participants (25%) chose the system in the *animation* condition as the one with the best performance, 6 participants (37.5%) selected the system in the *no-animation* condition as the one with the best performance, while the remaining 6 (37.5%) indicated that both systems had the same performance level. Only four participants answered the *non-reward-based* question differently than the *reward-based* question. The first participant changed from "both systems" to "animation", the second participant changed from "both systems" to

"no-animation". The third participant changed from "animation" to "both systems". Finally, the last participant change from "no-animation" to "both systems". With these changes the amount of participant that select one system in for the *reward-based* question was similar to the *non-reward-based* question. These results are illustrated in Figure 6.6. The participant who reported knowing Filipino was in this group, and she selected the system in the *animation* condition as the one having the best performance for the *reward-based* question, while she indicated that both programs had the same performance for the *non-reward-based* question.

6.2.2.2 Reasons for choosing one system over the other – reward-based question.

We categorised participants' comments into themes based on the reasons why they chose a particular HWR as the one majority of the participants will choose as the one with the best performance. Each response was associated with one or two themes, with six themes found in total: *number of errors, generic, animation, type of errors,* and *others' opinion.* The five themes are the same as in Study 7. Figure 6.7 illustrates the frequencies of these themes, also classified on the reward-based question (*animation* or *no-animation*) for the **9-errors group**, while Figure 6.8 illustrates the frequencies for the **10-errors group**. The participant who reported knowing Filipino was in the 10-errors group, and her answer was associated with the theme *type of errors.*

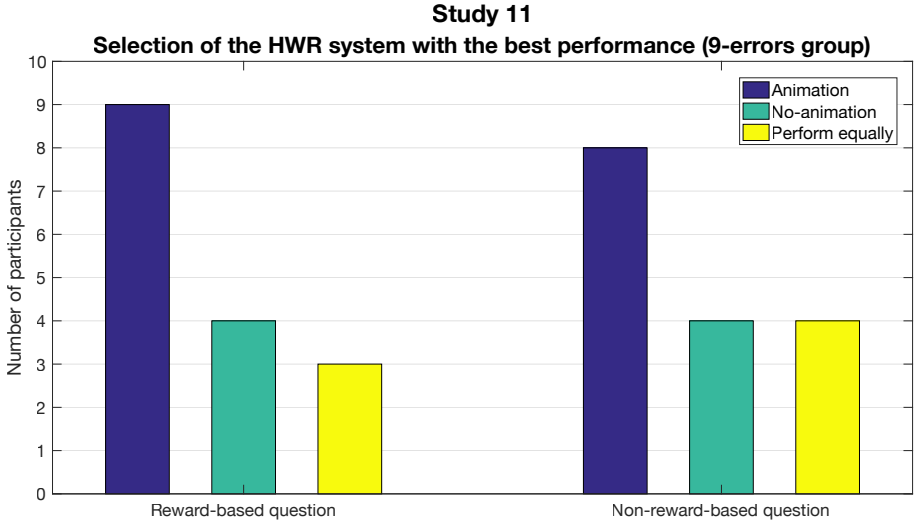

Figure 6.5: Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 11 for the 9-errors group. Number of participants on the y-axis.

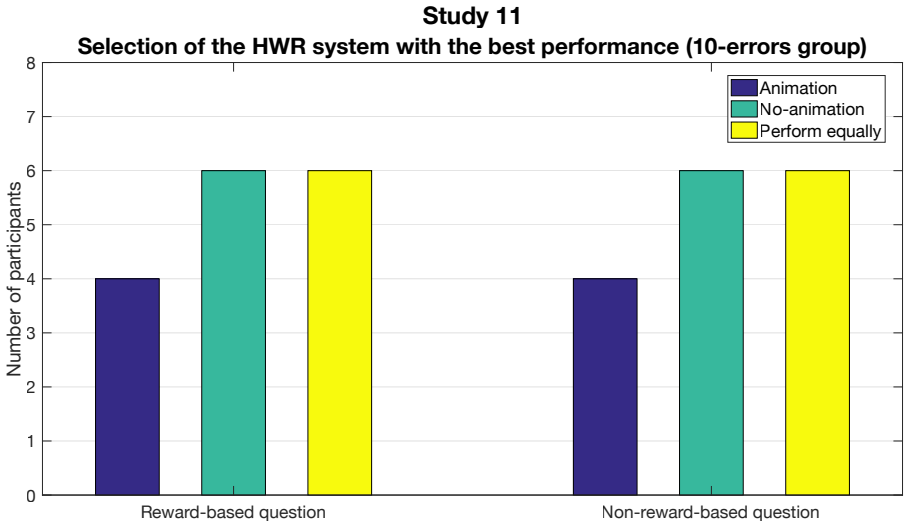

Figure 6.6: Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 11 for the 10-errors group. Number of participants on the y-axis.

6.2.2.3 Reasons for choosing one system over the other – non-reward-based question.

We grouped participants' responses into themes based on the reasons why they selected a particular HWR as the one with the best performance. Each response was associated with one or two themes, with four themes found in total: *number of errors, generic, speed* and *type of errors*. The themes categorised participants' comments as we did in

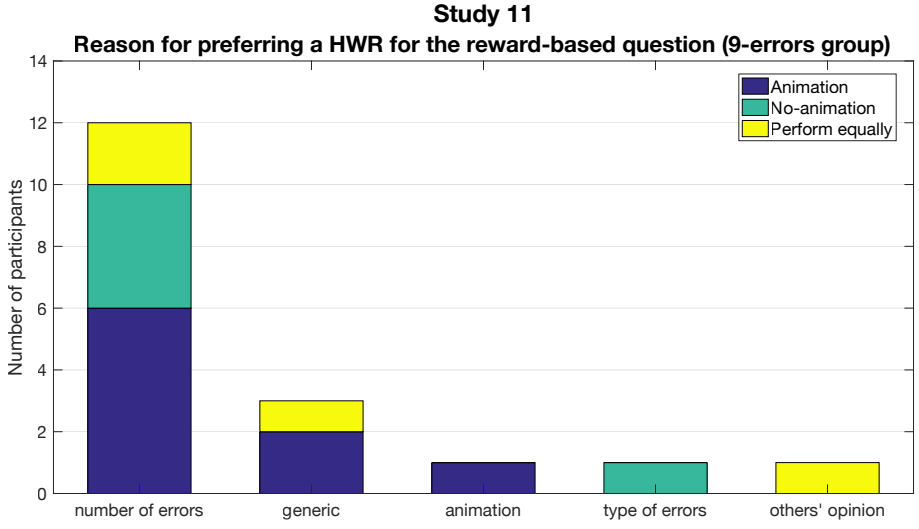

Figure 6.7: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question for the 9-errors group. Number of participants on the y-axis.

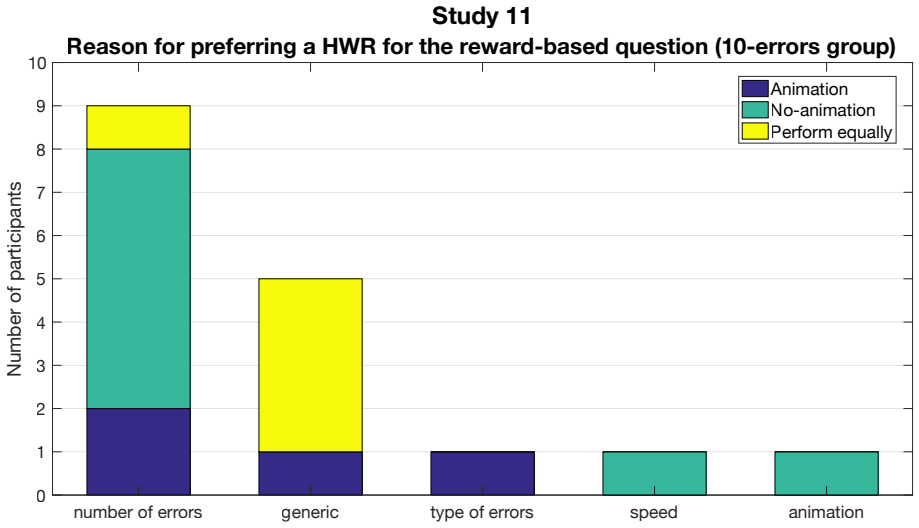

Figure 6.8: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question for the 10-errors group. Number of participants on the y-axis.

the previous studies. Figure 6.9 illustrates the frequencies of these themes, also classified on the individual preference (*animation* or *no-animation*) for the **9-errors group**, while Figure 6.10 illustrates the frequencies for the **10-errors group**. Similar to the previous subsection, the answer of the participant who reported knowing Filipino was associated with the theme *type of errors*.

Only one participant in the group 9-errors changed her selection from *animation* condition to "both systems" performed equally. The reason behind why the participant

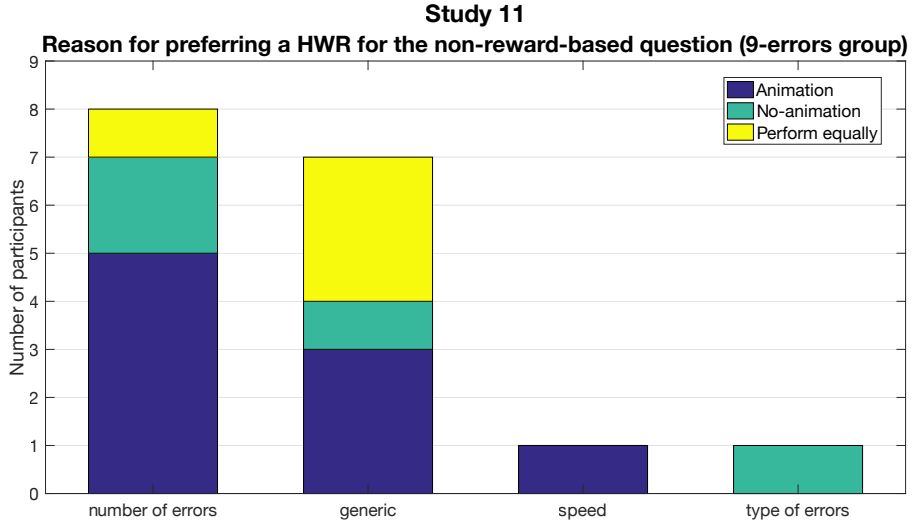

Figure 6.9: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the non-reward-based question for the 9-errors group. Number of participants on the y-axis.

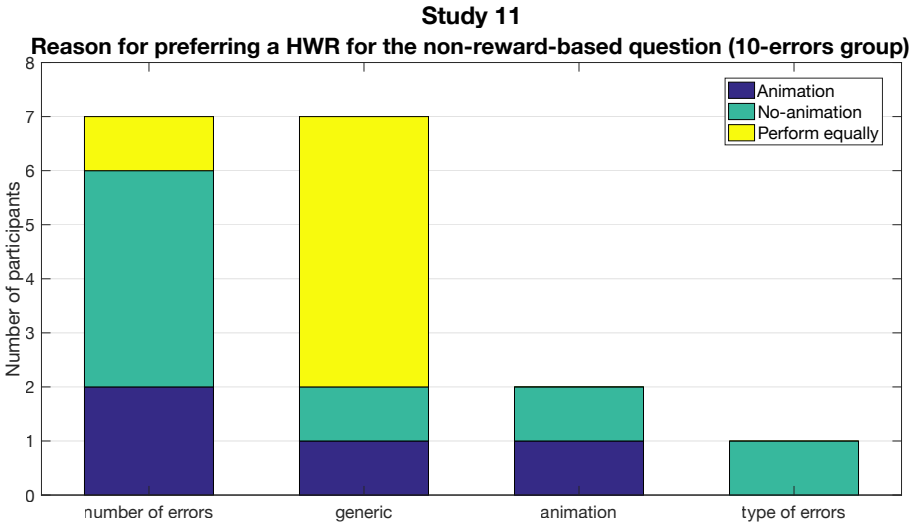

Figure 6.10: Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the non-reward-based question for the 10-errors group. Number of participants on the y-axis.

changed her choice is due to the fact the participant could see how the system worked in its task.

For the purpose of the analysis, we report each group corresponding to a number of errors separately. This applies both to the quantitative and the qualitative analysis.

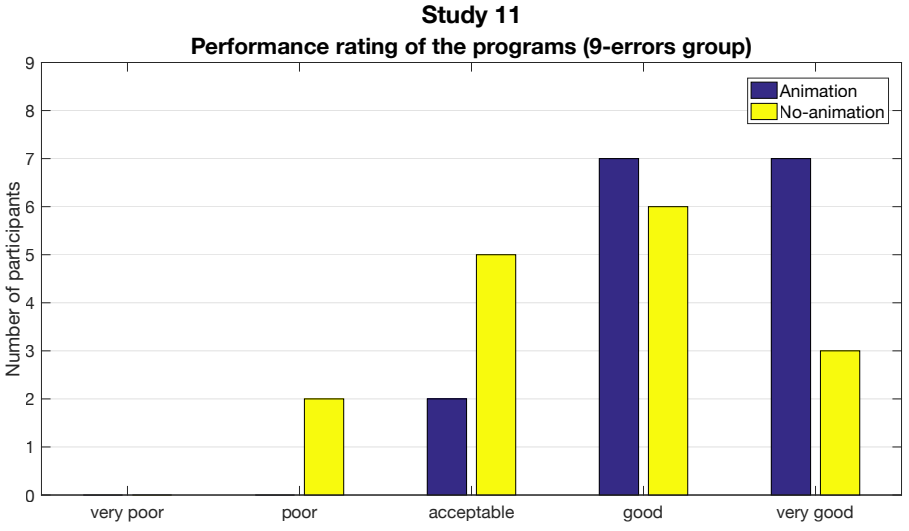

Figure 6.11: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions for the 9-errors group. Number of participants on the y-axis.

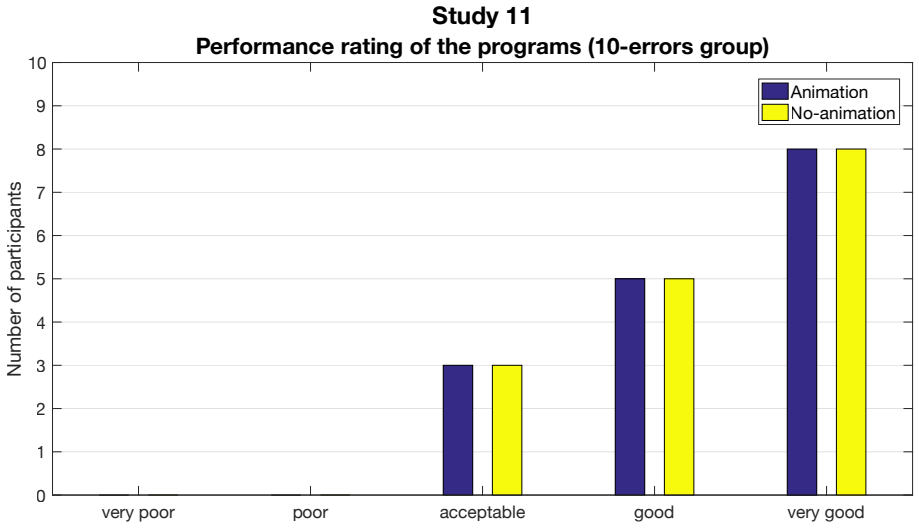

Figure 6.12: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions for the 10-errors group. Number of participants on the y-axis.

6.2.2.4 Performance ratings

9-errors group. A Wilcoxon Signed-ranks Test revealed that the performance evaluation was higher in the *animation* condition ($Mdn = 4$) than in the *no-animation* condition ($Mdn = 4$), ($Z = 2.183, p < .05, r = 0.39$). Figure 6.11 shows participant's evaluation of the performance of the systems.

10-errors group. A Wilcoxon Signed-ranks test did not reveal statistical significance suggesting the positive effect of animation was cancelled ($Z = 0, p = 1, r = 0$). Figure 6.12 shows participant's evaluation of the performance of the systems.

6.2.3 Discussion

The results of Study 11 indicate that the effect of *animation cues* on participants' perception of the system performance holds, to some extent, even when comparing two systems that have different performance levels. In particular, within the 9-errors group, most participants selected the system in the *animation* condition, even when it produced one additional mistake compared to the system in the *no-animation* condition (corresponding to a performance degradation of 12.5%). When the difference in number of errors produced by the two systems becomes 2 (the 10-errors group, which corresponds to a performance degradation of 25%), the animation system is no longer selected as the one with the best performance by the majority of participants, but only by 4 participants (25%). However, even in the 10-errors group, 6 participants (37.5%) suggested that the two systems have the same performance, and that's as many as those who correctly

selected the system in the *no-animation* condition as the one with the best performance. This finding is reinforced by the qualitative data; this shows that in both the 9-errors group and the 10-errors group, some participants suggested that there are fewer errors in the *animation* condition compared to the *no-animation* condition.

More in general, from this study, we can learn that the positive effect of *animation cues* can persist even when a system's performance is degraded. In contrast, Kim and Hinds (2006) showed that people who worked in a cooperative environment, blame others when a robot has a problem, and this one delivers an explanation of its failure. In more detail, people did not consider that the problem was a malfunction of the robot. Instead, they considered that the problems were due to interactions with other people. In contrast, our findings show that the *animation cues* tend to hide a possible malfunction of the system.

Finally, the answers submitted by the participant who reported knowing the Filipino language suggest once again that the knowledge of the language did not influence the behaviour in our study.

6.3 Study 12 – Variation of speed of animation cues vs. no-animation

The aim of this study is to test the research question analyse the effect of *animation cues* when the speed of the animation varies from fast motion to slow motion, vice-versa, and a raise in the speed in the middle of the animation. The aim to analyse the variation of the speed in the animation is to test our research question "Does varying the speed of the animations would "break the illusion" created by the *animation cues*?" As such, three animations were designed for the study, and we tested each animation against the *no-motion* condition we used in the previous studies.

6.3.1 Method

6.3.1.1 Study Design

For the design of Study 12, we replaced the animation used in previous studies with three new animations that we designed to show three different variations in the speed of the animation. We refer to the *animation* condition used in Study 5 as a *fast-to-slow animation* (*fast-slow*) condition for the first animation. The second animation that we designed was referred as a *slow-to-fast animation* (*slow-fast*) condition. Finally, the last animation was referred as a *slow-fast-slow animation* (*slow-fast-slow*) condition. We used a fully counterbalanced, between-participants design. One-third of the participants interacted with *fast-slow* and *no-motion* conditions. The second third visualise

with *slow-fast* and *no-motion* conditions. Finally, the last third of the participants interacted with *slow-fast-slow* and *no-motion* conditions.Again, we were especially careful in keeping a number of variables that could affect participants' perception of the system performance constant. These variables were the same as those listed in Study 5.

6.3.1.2 Participants

We recruited 48 participants through MTurk, using the same restrictions for recruitment in the previous MTurk studies. Their feedback nor their results did not suggest that they were not committed to the task. The age of participants ranged from 19 to 63 ($MD = 31.5, SD = 10.63$), where 24 were males (50%) and 24 were females (50%). Most of the participants were American, only 5 participants were of a different nationality: British, Polish, German and Indian. The education level of the participants ranged from primary school level to master's degree or equivalent level. Overall 22 of them had a university degree level, 18 a secondary school level, six master's degree level, and two a primary school level. The compensation for taking part in the study was $1.17 US dollars, which was calculated from a minimum wage of $10 per hour. Additionally, 27 participants received a bonus of $1.17 for the consensus-oriented reward mechanism design that we used.

6.3.1.3 Equipment

We used the same web page as in Study 5. However, the system for the *animation* condition was modified to show the new three animations that were designed.

6.3.1.4 Procedure

For this study, we follow the procedure in Study 5. However, the animation in Study 5 was changed for the new three animations that we designed to test the effect of *animation cues* when the speed of the animation variate. For the first animation, the speed of highlighting and appearing one by one of the first half of the words was similar to animation on Study 5. For the second half of the words, the speed of highlighting and appearing of the words was reduced by half[1]. This first animation's speed varied from fast to slow, giving the impression that the system slows down its speed. This animation was referred as a *fast-slow* condition. For the second animation, the speed for the first half of the words was decreased in half in comparison with the animation we used in Study 5 and the second half of the words kept the original speed[2]. This animation's speed varied from slow to fast giving an impression that the system suddenly started to

[1] https://vimeo.com/210298361.
[2] https://vimeo.com/210298382.

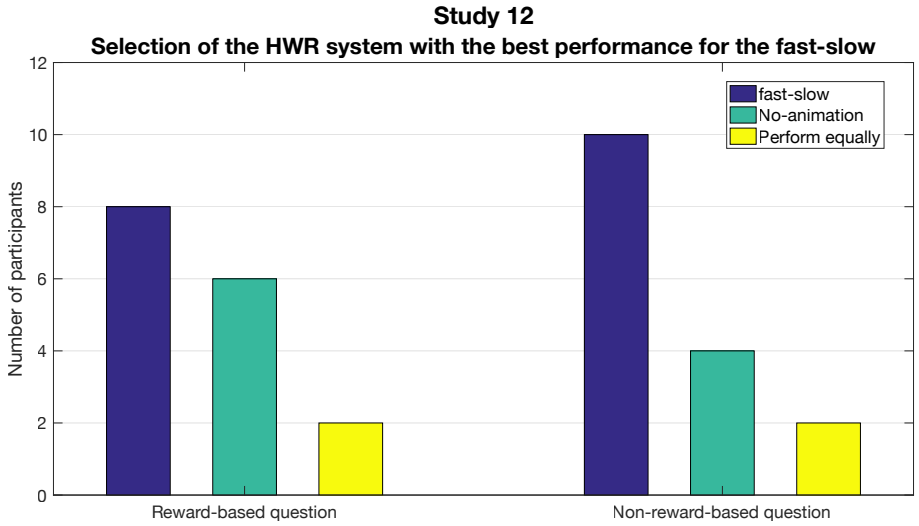

Figure 6.13: Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 12 for the *fast-slow* and *no-animation* condition. Number of participants on the y-axis.

work faster. The animation was referred as a *slow-fast* condition the second animation. Finally, the third animation the total number of words animated was split into thirds. The first and the last third of the animation, the speed was decreased in half as we did for the animation one and second third its speed was similar to Study 5[3]. The aim to keep the same speed for the second third of the words was to create the effect that the animation had a peak on the speed to process its task at the middle of it. This last animation we referred as a *slow-fast-slow* condition.

6.3.2 Results

6.3.2.1 Selection of the HWR system with the best performance

Fast-slow condition. For the *fast-slow* condition, 8 of the 16 participants (50%) selected the HWR as the one with the best performance. In addition, six participants selected the *no-animation* HWR as the one with the best performance, and two participants considered that both systems performed equally. For these conditions, two participants changed their selection in the non-reward-based question from "*no-animation*" condition to "*fast-slow*" condition. These results are illustrated in Figure 6.13.

Slow-fast condition. For the *slow-fast* condition, 10 of the 16 participants (63%) selected the HWR as the one with the best performance. In addition, six participants selected the *no-animation* HWR as the one with the best performance. For these conditions, four participants changed their selection in the non-reward-based question. In

[3] https://vimeo.com/210298397.

more detail, two participants changed from "*no-animation*" condition to "*slow-fast*" condition, one participant from "*slow-fast*" to "*no-animation*" condition, and "*no-animation*" condition to "*both program*" perform equally. These results are illustrated in Figure 6.14.

slow-fast-slow condition. For the *slow-fast-slow* condition, 9 of the 16 participants (56%) selected the HWR as the one with the best performance. In addition, five participants selected the *no-animation* HWR as the one with the best performance, and two participants considered that both systems performed equally. For these conditions, four participants changed their selection in the non-reward-based question. In more detail, two participants changed from "*both program*" perform equally to "*fast-slow*" condition, one from "*no-animation*" condition to "*both programs*" perform equally and one changed from "*no-animation*" condition to "*slow-fast-slow*" condition. These results are illustrated in Figure 6.15.

6.3.2.2 Reasons for choosing one HWR system over the other for the reward-based question.

Participants' responses to the question about they chose a particular HWR as the one majority of the participants will choose as the one performing best were categorised through thematic analysis. Each response was associated with one or two themes, with five themes found in total: *number of errors, generic, animation, typos, speed* and *others' opinion*. The five themes are the same as in Study 7. Figure 6.16 illustrates the

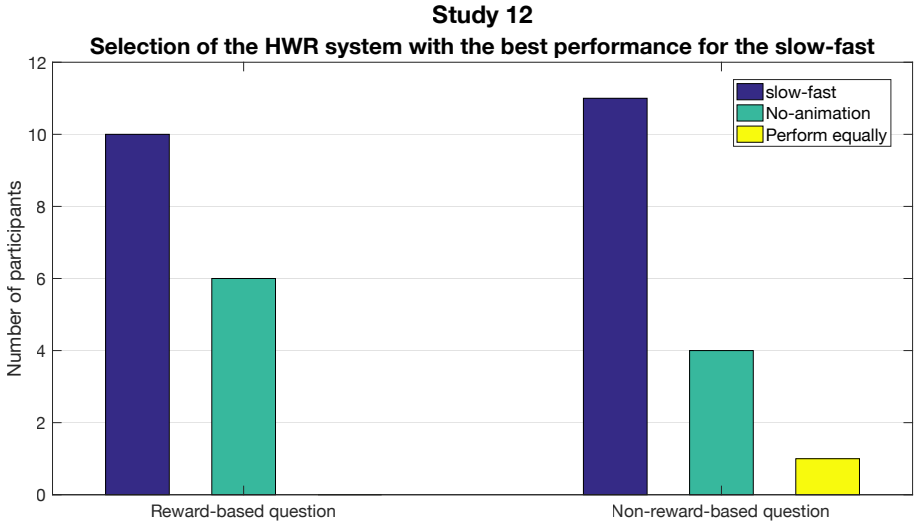

Figure 6.14: Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 12 for the *slow-fast* and *no-animation* condition. Number of participants on the y-axis.

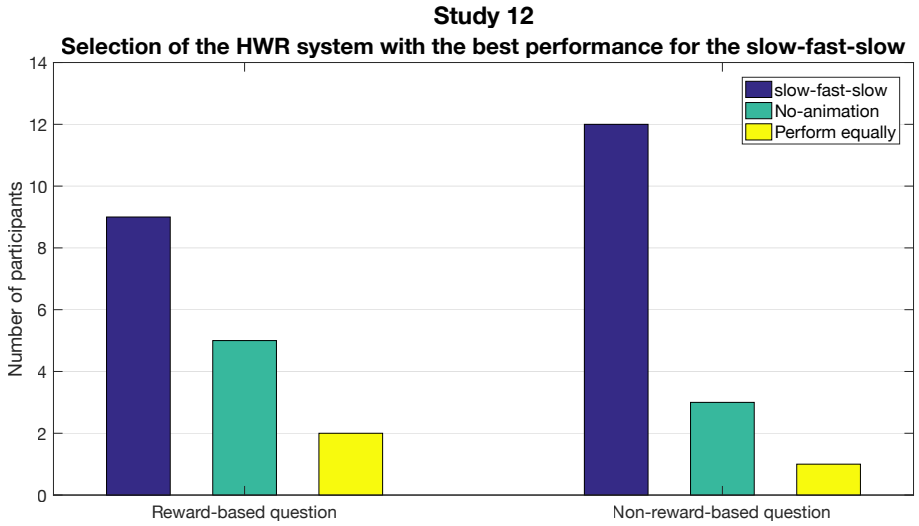

Figure 6.15: Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 12 for the *slow-fast-slow* and *no-animation* condition. Number of participants on the y-axis.

frequencies of these themes, also classified on the *reward-based* question *fast-slow* or *no-animation*, while Figure 6.17 illustrates the frequencies for the *slow-fast* or *no-animation*. Finally, Figure 6.18 illustrates the frequencies for the *slow-fast-slow* or *no-animation*.

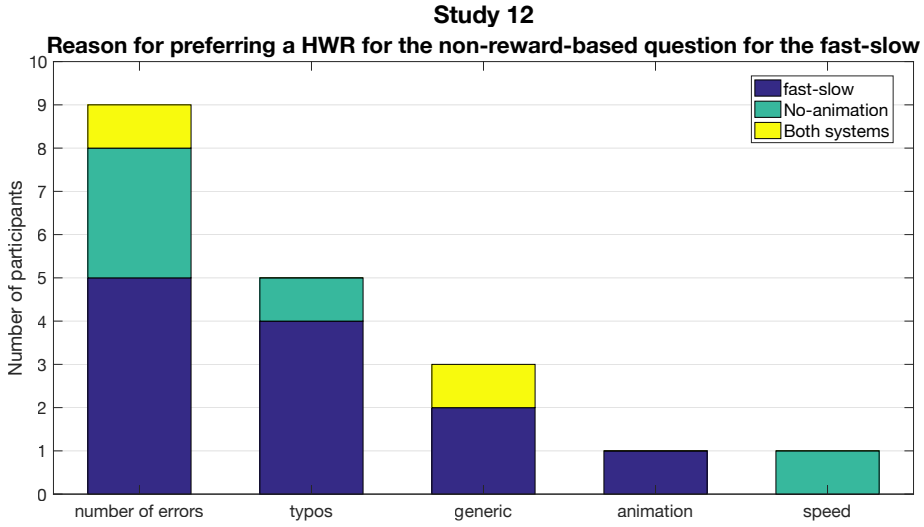

Figure 6.16: Reasons expressed by participants for preferring a HWR in the *fast-slow* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

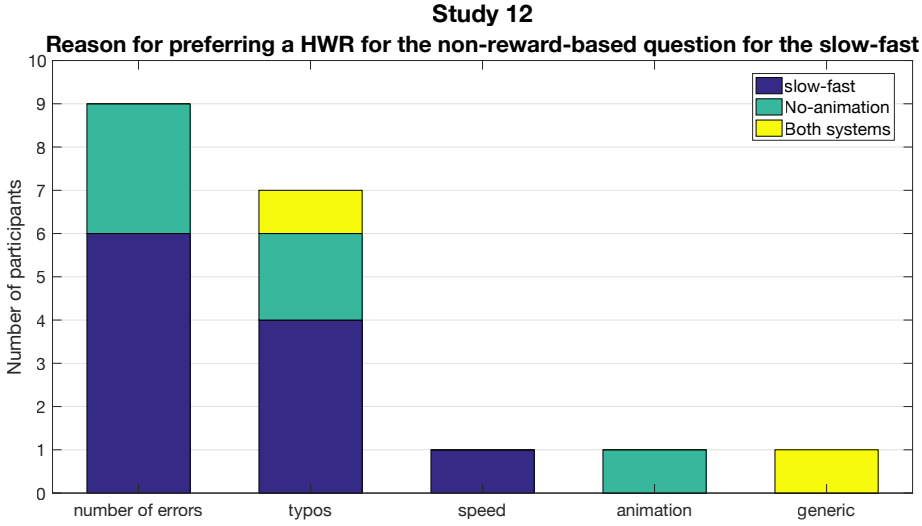

Figure 6.17: Reasons expressed by participants for preferring a HWR in the *slow-fast* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

6.3.2.3 Reasons for choosing one HWR system over the other for the non-reward-based question.

We themed participants' responses based on the reasons why they selected a particular HWR as the one with the best performance. Each response was associated with one or two themes, with five themes found in total: *number of errors, generic, speed, animation* and *typos*. The themes categorised participants' comments as we did in the

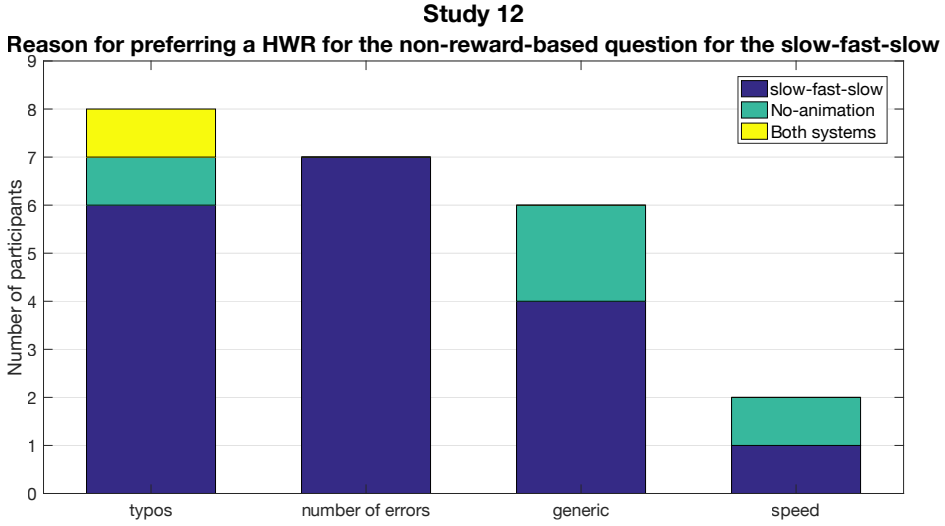

Figure 6.18: Reasons expressed by participants for preferring a HWR in the *slow-fast-slow* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

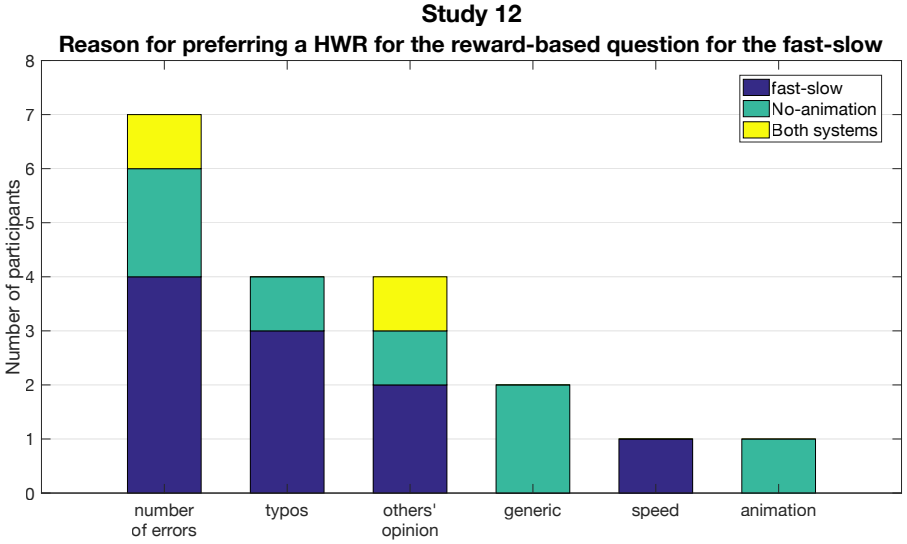

Figure 6.19: Reasons expressed by participants for preferring a HWR in the *fast-slow* or *no-animation* condition for the *non-reward-based* question. Number of participants on the y-axis.

previous studies. Figure 6.19 illustrates the frequencies of these themes, also classified on the *non-reward-based* question *fast-slow* or *no-animation*, while Figure 6.20 illustrates the frequencies for the *slow-fast* or *no-animation*. Finally, Figure 6.21 illustrates the frequencies for the *slow-fast-slow* or *no-animation*.

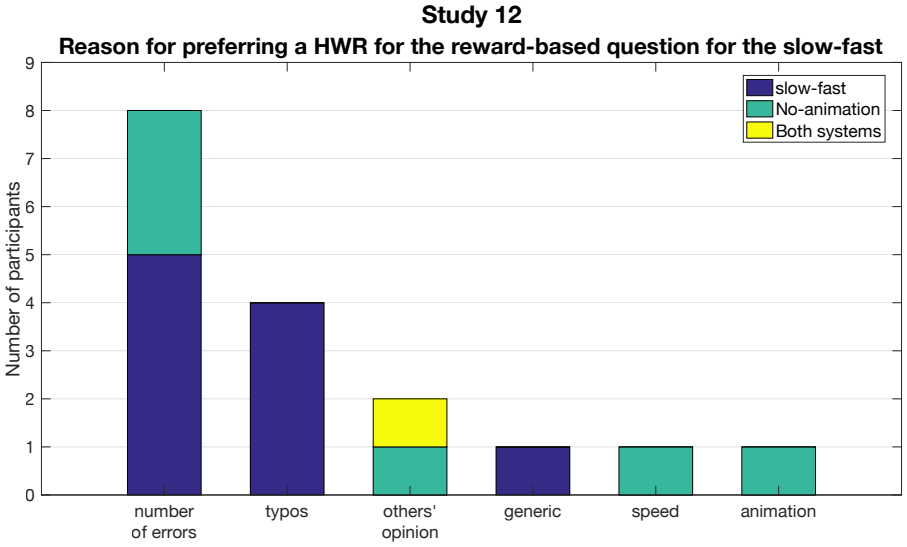

Figure 6.20: Reasons expressed by participants for preferring a HWR in the *slow-fast* or *no-animation* condition for the *non-reward-based* question. Number of participants on the y-axis.

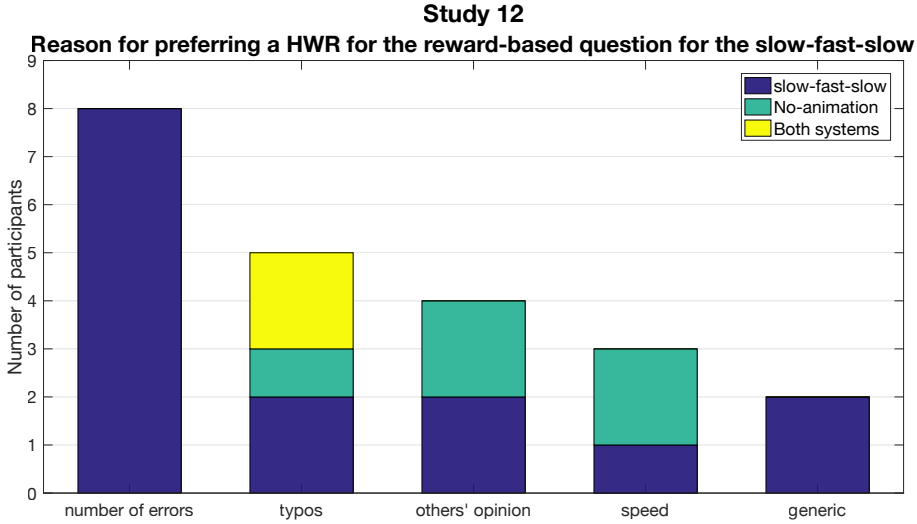

Figure 6.21: Reasons expressed by participants for preferring a HWR in the *slow-fast-slow* or *no-animation* condition for the *non-reward-based* question. Number of participants on the y-axis.

6.3.2.4 Performance evaluation of the HWR

***Fast-slow* condition.** A Wilcoxon Signed-ranks Test revealed that the performance evaluation was higher in the *fast-slow* condition ($Mdn = 4.5$) than in the *no-animation* condition ($Mdn = 3$), ($Z = 2.39, p < 0.05, r = 0.42$). Figure 6.22 shows participant's evaluation of the performance of the systems.

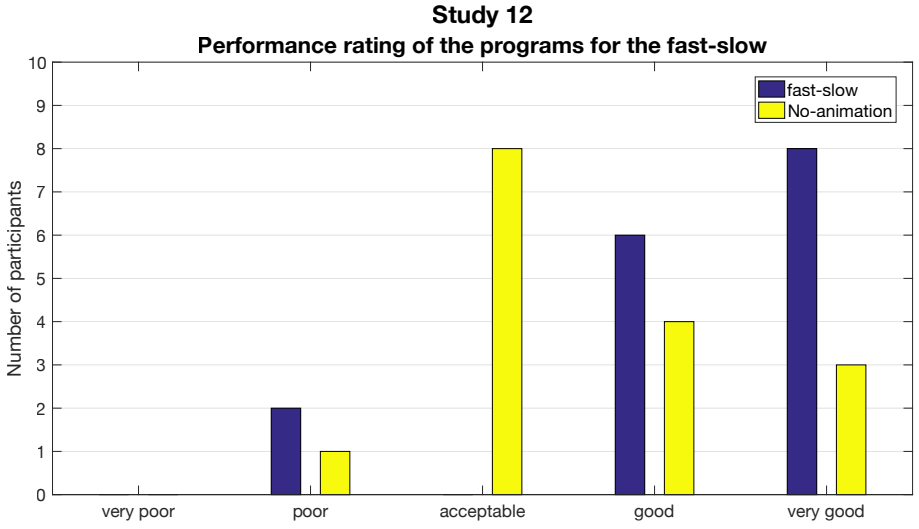

Figure 6.22: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

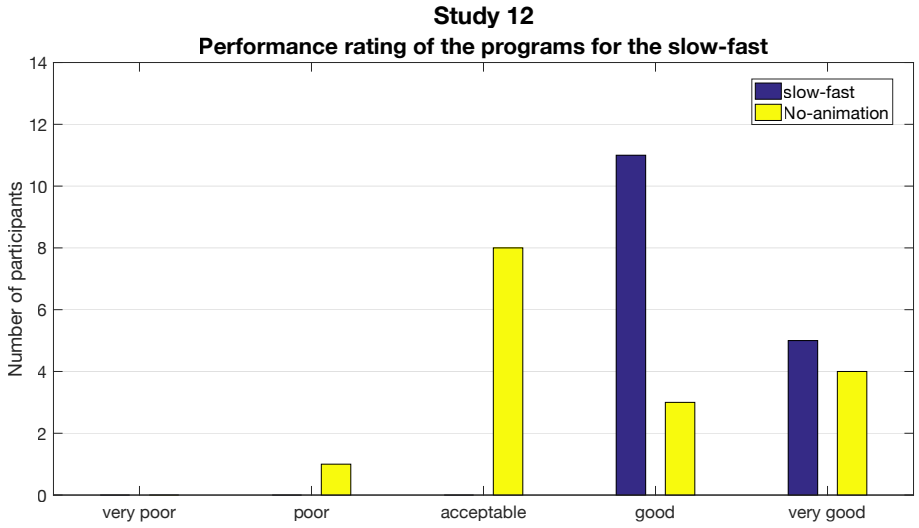

Figure 6.23: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

***Slow-fast* condition.** A Wilcoxon Signed-ranks Test revealed that the performance evaluation was higher in the *slow-fast* condition ($Mdn = 4$) than in the *no-animation* condition ($Mdn = 3$), ($Z = 2.30, p < 0.05, r = 0.41$). Figure 6.23 shows participant's evaluation of the performance of the systems.

***Slow-fast-slow* condition.** A Wilcoxon Signed-ranks Test revealed that the performance evaluation was higher in the *slow-fast-slow* condition ($Mdn = 4$) than in the *no-animation* condition ($Mdn = 3.5$), ($Z = 2.68, p < 0.01, r = 0.47$). Figure 6.24 shows participant's evaluation of the performance of the systems.

6.3.3 Discussion

The findings of this study clearly show that the variation in the animation speed did not cancel the positive effect of *animation cues* of how people perceive systems' performance. However, we found that in the *fast-slow* condition only half of the participants selected the system with the *animation cue* to be the one with the best performance. Additionally, in the *slow-fast* condition 63% of the participants selected the system with the *animation cue* to be the one with the best performance. These results are aligned to those Tremoulet and Feldman (2000) found in their work. Tremoulet and Feldman [Idib.] found that objects in a screen that accelerate yield greater animacy rating than ones that decelerate.

However, from the results of Study 4 and 5, we found that more people considered that the system with the *animation cues* was the one with the best performance and in this study fewer people selected the system that showed *animation cue*. Thus, we consider

that designers need to keep the same speed for animation timelapse to keep homogeneity in the animation and increase the rate acceptance of the people.

6.4 Summary

From Study 10 (N=48) and Study 12 (N=48), we found that varying other dimensions of *animation cues*, such as the amount of detail and speed, did not affect how people perceive the performance of screen-based systems. However, we consider that *animation cues* need to be shown on both the input and output parts of the UI. Moreover, animation's speed needs to keep homogeneity without showing any acceleration or deceleration. In contrast, our results in Study 11 (N=32) suggest that the positive effect of *animation cues* can be cancelled once the performance of the system that shows the *animation cue* has a degradation of 25%.

These results have a design implication in how the *animation cues* need to be designed to increase the acceptance of the users as *visual feedback*. However, the results suggest that any variation in the amount of detail of speed in the animations will not cancel the positive effect of the *animation cues* on how people perceive the performance of screen-based systems.

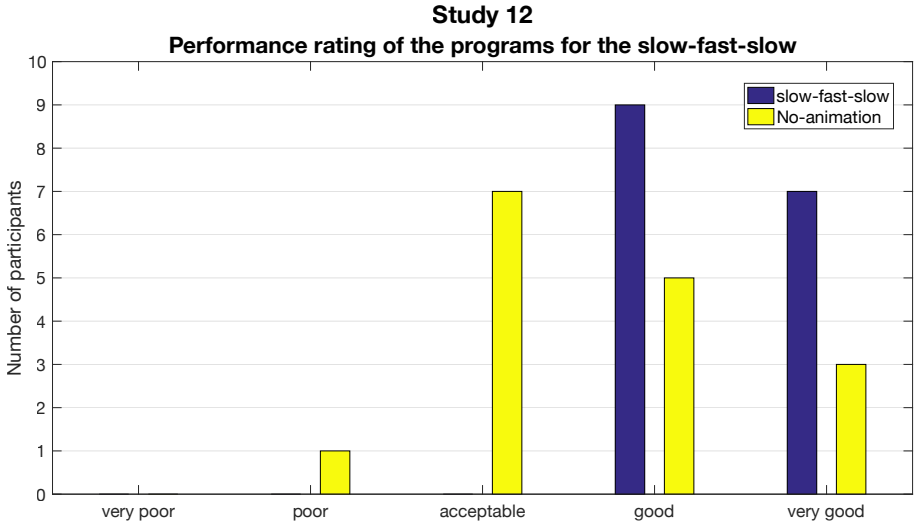

Figure 6.24: Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

Chapter 7

General Discussion and Conclusions

In this research, we have explored the effect of *visual cues* on how people perceive the performance of smart systems (i.e. robot and screen-based systems). We ran 12 studies to test our research questions and observe the effect of *physical motion cues* on vacuum cleaning robot systems and *animation cues* on screen-based systems. In more detail, the first three studies were designed to observe whether *physical motion cues* have an effect on people's perception of robot systems. Additionally, the effect was compared with the effect of a different modality that designers could implement in their design (i.e. video-based notification). Moreover, the rest studies were designed to test the effect of *animation cues* on screen-based systems. In addition, we focused on understanding why this effect was observed because *animation cues* are not intrinsic to the systems as *physical motion cues* are for the robot systems.

In more detail, Study 1 was designed to analyse whether *physical motion cues* can influence people's perception of the performance of vacuum cleaning robots. Our participants compared the performance of two robot system, one robot show *physical motion cues* as a feedback and the other robot only delivered a text notification. Participants' evaluation suggests that the robot that shows *physical motion cues* was considered to be the one with the best performance. Additionally, we designed Study 2 to analyse whether *physical motion cues* are more effective than *video-based cues* at influencing how people evaluate a vacuum cleaning robot that shows such cues. For this study, participants compared the two robots, similar to what the participants in Study 1 did. However, in contrast to Study 1, the text notification sent by the robot in one condition was replaced with a video notification. Study 2 findings suggest that indeed *physical motion cues* are more effective than *video-based cues*. These two studies confirm that showing *physical motion cues* is an effective approach to affect how people perceive robot systems' performance. Finally, we ran Study 3 to analyse whether *video-based cues* can

influence people's perception of vacuum cleaning robots' performance. Thus, we remove the *physical motion cue* from Study 2, and we left the robot with only text notification to confirm our research question. However, we observed that *video-based cues* are not effective to change people's perception, even though, people claim that they prefer to have *video-based cues* as a feedback to assure that the robots work in their task, also for practicality.

Based on the findings we observed in Studies 1, 2, and 3, we ran three more studies to analyse whether *animation cues* have an effect on screen-based systems, even if in this type of system the animations shown as a feedback are not intrinsic to their functionality. Study 4, was designed to compare two HWR systems to compare how participants evaluated systems' performance when they receive *animation cues* as feedback in one of the systems. The results of Study 4 suggest that *animation cues* also affect how people perceive screen-based systems' performance. These results are an extension of those whom we found in the first three studies. Additionally, Study 5 was designed to observe whether this effect is also seen in a non-controlled environment. As such, we ran this study in a crowdsourcing platform (MTurk). The results of this study show that the implementation of *animation cues* as feedback can be implemented in a non-controlled environment. Moreover, we found that a crowdsourcing platform is a tool that we can use to gather a variety of participants for future studies. Finally, Study 6 was designed to observe whether this effect can also be observed for another type of screen-based system (i.e. part-of-speech systems). The findings of Study 6 show that *animation cues* also have an effect on people's perception. As such, we suggest that the implementation of *animation cues* as feedback can be applied to screen-based systems.

A natural step after we confirm that *animation cues* have an effect on people's perception is to understand what kind of characteristics should animations have to affect how people perceive screen-based systems. Hence, Study 7, 8, and 9 were designed to analyse whether the effect is because the animations being human-like or the animations match people's mental model. The results of these studies show that the animation we implemented for Study 4 and 5 matches with people's mental model. In more detail, in Study 9 we observed that participants that received an explanation of how the systems worked and matched with the *animation cues* they visualised, perceived that the systems performed better than the participants who received an explanation that mismatch how the systems worked.

In addition, building on the results of Studies 4, 5, and 6, we analysed the effect of varying other dimensions of the *animation cues* (i.e. the amount of detail of the animation, number of errors in the outcome, and speed of the animation). The results of Study 10 showed that the amount of detail of the *animation cues* did not produce the same effect we found in Studies 4, 5 and 6. Moreover, Study 11 findings show that the effect observed persisted when the number of errors increases in 12.5% for the system that shows the *animation cue*. Finally, we ran Study 12 to analyse whether the varying of

the *animation cues'* speed affects the effect we found in previous studies. In particular, the findings of this study show that the effect persist even the variations in the speed.

Furthermore, in this chapter, we discuss the implications for the design that the implementation of *visual cues* has on smart systems. As such, we propose scenarios where designers can integrate these cues in their design or avoid their application to eliminate bias that can affect the interaction between their systems with their end users. Finally, we want to motivate researchers to investigate further other implications or effects that *visual cues* can have in other systems or variation in the design.

7.1 General Discussion

The results of the 12 studies show clearly that the *visual cues*, as a form of *visual feedback*, can have a considerable effect on how people perceive the performance of smart systems. This was true for both a physically embodied system (vacuum cleaning robot) and an autonomous screen-based system, for which results were displayed on a standard computer screen. In particular, people perceive that such systems performed better when *visual cues* are displayed than when it is not. These findings confirm the first and fourth research question that *visual cues* influence people's perception of the performance of smart systems.

Compared to prior work (Hoffman and Vanunu (2013); Vermeulen et al. (2013)), the systems in our study simplistically displayed motion without complicating systems' structure. The motion presented by the systems is intrinsic to the system or is an animation that gives an impression to the user of how the screen-based system processes its task. The aim of this interaction is not to modify the structure of the system, removing the need to add actuators that trigger *visual cues*. Indeed, in the case of the Roomba, the motion cue we studied is simply part of the standard operation of the robot. The motion presented on the HWR and POS systems is a simple animation that emulates how these systems work.

From Studies 2 and 3, we learnt that showing *video-based cues* for robots is not an effective visualisation that can affect people's perception of how the system performs its task. However, the results show that people like to have a visual notification as a feedback. As we mentioned above, designers need to keep in mind that if they are looking to change people's perception, it is necessary to show *physical motion cues*. To this end, additional measures may need to be put in place to drive the user's attention to the motion. For example, presence or location sensing (including e.g. smartphone apps to detect the user's location) may be employed to activate the system when users are physically close to them, or on their trajectory home, leveraging prior work on pattern recognition on GPS traces (Horvitz and Krumm (2012)).

The results of Study 4, 5, and 6, revealed that *animation cues* could have an effect on the perceived performance of screen-based systems as in physically embodied systems (i.e. robots). In particular, our participants rated the system which displayed an animation as performed better than an identical system not showing any animation. In more detail, the results of Study 5 confirmed that in a less controlled environment people's perception was affected after they saw the *animation cues*. Moreover, Study 6 revealed that the effect of *animation cues* are not restricted only to Optical Character Recognition systems (i.e. Handwritten Recognition system) but also to Natural Language Processing system (i.e. Part-of-Speech system). These results suggest that *animation cues* can be applied in a wide range of screen-based systems. Additionally, these results extend the findings from studies of robot systems to screen-based systems. Such an extension makes these results applicable to a wider range of systems, including mobile and web applications such as translators, recommender systems or even chatbots in e.g. automated customer support.

After having observed the strong effect of animations on the perceived performance of smart systems, we designed a number of follow-up studies aimed at explaining the cause of such an effect. Study 7 ruled out the possibility that the effect is because the animation making the systems look human-like (and hence as smart as a human). Instead, the positive effect of animations appears to persist even when the animation is clearly not human-like, although the effect is not as strong anymore. Study 9 investigated the relationship between mental models and the animation effect, bringing to light that the effect noticed in Study 4, 5 and 6 only occurs when the displayed animation is similar to the user's mental model of how the system works.

Once we found the reason behind the effect of *animation cues* has on people's perception, we focus on observing whether the detail of the animation implemented can impact the effectiveness of the cues. As such, Study 10 validated that higher detail of animations can further influence people's perception of smart systems. Moreover, Study 11, we found that the effect of the *animation cues* persisted even if the performance of a smart system has decreased minimally. Hence, as designers, we should be aware that this bias can be both seen as positive or negative, depending on the context and implementation of the system. In the next section, we discuss this issue and present a number of design guidelines. Finally, in Study 12 was designed to observe the impact of speed variation in the time-lapse of the animation in the effect of *animation cues*. Thus, three animations were designed to test the change of the speed. The results of Study 12 suggest that the effect of *animation cues* persist even if the speed of time-lapse of the animation varies.

7.2 Implication for design

Visual cues can potentially be applied to a wide range of devices, such as robots and screen-based systems. In the domestic context, smart appliances such as vacuum cleaning robots, washing machines (e.g., seeing the spin cycle confirms the clothes will be clean and dry) and dishwashers (e.g., hearing the dishwasher rinse and shut down confirms all dishes have been cleaned) can be timed according to GPS traces such that when they detect (or predict) that the owners are nearby, they will finish their cycles. Moreover, we encourage designers to implement this approach given the simplicity of its implementation on top of existing designs.

A similar approach can also be used for vending machines that prepare food, such as coffee[1], chips[2], and pizza[3]. On the other hand, the interaction can be implemented on screen-based system, such as websites that show meaningful animations (e.g., suggested lists of products being processed or emailing confirmations) when loading content or when processing information takes place on the server (in more elaborate ways than simple ajax loaders gifs such as a spinning pinwheel). However, our approach for screen-based systems requires that designers keep in mind the existence of some factors that can cancel the positive effect on people's perception. As such, we present in the following subsection recommendations that designers should follow to guarantee the effectiveness of our approach.

7.2.1 Design of Screen-based Systems

Our studies bear implications for the design of screen-based systems, and in particular the design of *visual feedback* around such systems. While the effect of changing a person's perception through an animation can be explicitly used to a designer's advantage, to make a screen-based system be considered more favourably by users, it is equally important to be aware of possible unintended biases. It may be detrimental to make a probabilistic system appear more accurate than it is. At the opposite end, including *animation cues* that are, unintentionally perhaps, at odds with how users perceive the system to work may diminish the users' confidence in the system. As findings from studies 4 to 6 showed, designers need to realise that people can perceive a system's performance better because of the implementation of *animation cues*. Furthermore, it is necessary to make the *animation cues* compatible with the users' mental model of how the system works. As such, designers can run interviews to understand how users believe their systems work. Furthermore, in the design of new systems, designers need to run pilot studies to analyse people's mental model of how they consider the systems work or perhaps provide explicit instructions that explain the system's operation in a

[1] http://www.nwglobalvending.co.uk/products-brands/vending/canto-touch.
[2] http://www.beyondte.com/.
[3] http://wonderpizzausa.com/.

way that is compatible with the *animation cues*. As Lim et al. (2009)'s findings suggest that users have a better understanding of a system's behaviour and a higher feeling of trust in it when it provides explanations.

Additionally, our findings from studies 4 to 6 suggest that users may be influenced by *animation cues* that lead them to overlook potential system failures. As such, this bias can cause significant harm when dealing with safety-critical systems, provoking users to be in dangerous situations. Hence, we propose that designers should avoid the implementation of this *visual feedback* in safety-critical systems to avoid influencing users' perception of systems' performance.

7.2.2 Limitations and Future Work

We believe that the effect we observed may have potential to influence people's inclination to adopt such systems, even though more research is required in such direction. These studies suggest that *visual cues* can influence how people perceive the performance of smart systems, such as vacuum robots and screen-based systems. However, we found some limitations in our experimentation, as such, we propose future work base on the limitations we identified.

We identified that the majority of the studies in this thesis are based on a binary comparison. This means that we only compared *visual cues* that only show the systems showing motion at the end of their task against systems that did not show any cue about their performance. Hence, we believe that another approach might be to evaluate the effect of *visual cues*. The first approach we propose for future studies can be made by multiple comparisons of the systems that show *visual cues* against a system that did not show any cue and a system with other modality as a feedback. The second approach is designing a between-subjects study where one-half of the participants will receive the *visual cue* treatment; the second half will receive the no *visual cue* treatment. This last approach can be useful for in the wild studies where sometimes it is problematic that participants can have multiple treatments.

Additionally, we only look at short term effect. While this is a needed contribution, future work should look into whether there are any long-term effects. As such, future researchers should run diary studies or pop-up studies to observe the effect of *visual cues* in everyday systems that participants will interact with them in their daily activities. Furthermore, in these studies, the systems will be showing the *visual cues* only one-half of the duration of the study and the second half the *visual cues* can be removed or vice-versa. The aim of this design is to analyse how people perceive systems performance when the *visual cues* disappear or appear during the study. However, we consider that this kind of studies can represent a challenge because a malfunction in the system or tracking participants to trigger the *visual cues* can represent a problem for the study and

the results. Even if is not plausible to run these studies, we already found that the effect on short term interactions. Hence, there are applications where this interaction could still be applied, such as those that are used only once or infrequently (e.g. seeing self-checkout machines processing users' purchases or an automatic water sprinklers being activated before people arrive home).

In the Roomba studies, we only focused on analysing *physical motion cues* that showed the robot docking once participants entered to evaluate the cleanliness of the carpets. However, more research needs to be conducted around other robots and tasks where researchers can run more studies to analyse the effect of *physical motion cues* on how people perceive the performance of robots. Moreover, researchers can observe whether the robots showing the *physical motion cues* at the beginning of the task can yield the same results we found in Studies 1 and 2.

Finally, in Chapter 4, 5, and 6, we found that *animation cues* affect how people perceive screen-based systems' performance and which characteristics these *animation cues* need to have. These studies can be run with an eye tracker to analyse which part of the screen people are watching when the *animation cue* is shown. This data can help researchers to understand better people's behaviour after they receive an *animation cue*. Moreover, researchers can analyse whether people pay more attention to the area where the errors are located or whether participants spend the same time on analysing every word.

7.3 Conclusion

In this thesis, we presented twelve studies through a lab study and through the use of a crowdsourcing platform. These studies were designed to explore whether *physical motion cues* and *animation cues* can change people's perception in the evaluation of smart systems and what characteristics the animations do need to have. Indeed, our results indicate that *physical motion cues* have a positive effect on people's perception of robot systems performance in comparison to *video-based* feedback. Furthermore, our results suggest that displaying a high detail of animations that match people's mental model as *animation cues* can influence people's perception of the performance of smart screen-based systems even if these systems have a minimal decrease in their performance.

In general, our results indicate that this modality has potential to improve user ratings, as the display of *visual feedback* can change how people perceive and evaluate the system's performance, whether it be *physical motion* for robots systems or *animation* for screen-based systems. We hope that the results presented in this thesis will stimulate designers to integrate physical motion or animations as a feedback of their systems, and researchers to explore this area further.

Appendix A

Ubicomp Paper

In this appendix, we add the paper we presented in the international conference Ubicomp' 16.

The potential of physical motion cues: Changing people's perception of robots' performance

Pedro Garcia Garcia, Enrico Costanza, Sarvapali D. Ramchurn and Jhim Kiel M. Verame
University of Southampton
Southampton, United Kingdom
{pgg1g14, e.costanza, sdr1, j.verame} @soton.ac.uk

ABSTRACT
Autonomous robotic systems can automatically perform actions on behalf of users in the domestic environment to help people in their daily activities. Such systems aim to reduce users' cognitive and physical workload, and improve well-being. While the benefits of these systems are clear, recent studies suggest that users may misconstrue their performance of tasks. We see an opportunity in designing interaction techniques that improve how users perceive the performance of such systems. We report two lab studies (N=16 each) designed to investigate whether showing physical motion, which is showing the process of a system through movement (that is intrinsic to the system's task), of an autonomous system as it completes its task, affects how users perceive its performance. To ensure our studies are ecologically valid and to motivate participants to provide thoughtful responses we adopted consensus-oriented financial incentives. Our results suggest that physical presence does yield higher performance ratings.

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

Author Keywords
Visualisation of Automation; Cognitive Psychology; Automated Systems; Robots; User Experience; Perception

INTRODUCTION

There is a growing number of systems able to *automatically* perform actions on behalf of users. Such systems are becoming increasingly widespread in the domestic environment to help people in their daily activities, such as water sprinkling[1] and vacuum cleaning. Moreover, as the domestic environment becomes increasingly instrumented with smart sensors through the Internet of Things (IoT), autonomous systems will be essential to manage the wealth of data such sensors generate, relieving their users of significant cognitive and physical workload involved in performing their daily activities.

While the potential benefits of automatic systems are clear, there are open questions around how users would perceive such systems and their operation [16]. Recently, researchers have investigated the usefulness of existing IoT products, such as the Nest thermostat [25] and vacuum cleaning robots (e.g. iRobot's Roomba) [20, 12]. Results from these studies suggest that because these systems operate autonomously, users do not normally attend to them while they undertake their tasks, so there could be a mismatch between the system performance and users' perception of it. As such, we see an opportunity in designing interaction techniques that may improve how users perceive the performance of automatic systems. In particular, we focus on *notifications*: notification systems are often necessary to alert users when the autonomous operation has been completed (given that users may not attend to it). Is it possible, then, to engineer notifications generated by autonomous system to influence users perception of the system performance?

Against this background, in this paper we report on two lab studies designed to investigate whether showing in person the *physical motion* of an autonomous system as it completes its task can affect how users perceive its performance. Physical motion refers to the robot's movement as it processes its task, hence the motion we refer to is intrinsic to the system's task (this is in contrast to "physical motion" as being independent of task execution). Consensus-oriented financial incentives were used to increase the ecological validity of the studies[4] and motivate our participants to provide thoughtful responses. In particular, the first user study (N=16) focused on comparing two situations: (i) a moving robot in the process of docking and (ii) a static robot that has already completed its task. The aim was to see whether motion can positively change people's perception of the performance of an autonomous robot. Our results demonstrate that this is indeed the case: our participants almost unanimously rated the performance of the moving robot higher compared to a non-moving robot. In the second user study (N=16), we instead focused on investigating whether seeing the motion in person or through a video feed makes a difference. The results suggest it does: physical presence yields higher performance ratings. Indeed, the findings

[1] http://tinyurl.com/kzk9uuf
[1] http://tinyurl.com/legw4zt

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
UbiComp '16, September 12 - 16, 2016, Heidelberg, Germany
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4461-6/16/09. . . $15.00
DOI: http://dx.doi.org/10.1145/2971648.2971697

of this paper provide implications for how the feedback of autonomous systems can be enhanced to support how people perceive the performance of such systems.

RELATED WORK

Our research aims to evaluate how visual cues can change people's perception about the performance of autonomous robots. As such, we are building upon prior research that has studied transparency for the intelligibility of robotic systems, perception of motion in robots and interactive artefacts, and perception of motion in robots through video and animation. We next elaborate on the literature in these three key areas in turn.

Transparency for the Intelligibility of Robots

Robots are expected to become part of people's everyday life in homes and public areas (e.g. in hotels, trade shows, workplaces, museums) [19, 11]. Therefore, robots should be transparent about their decisions and actions so that people would feel that they understand their behaviour [13].

Kim et al. [13] examined whether different levels of transparency have an effect on people's judgement of blaming an autonomous robot or someone else at the moment the robot presents an unexpected behaviour in a cooperative scenario. In more detail, the robot delivered assemblies of toy pieces that participants place in a tray that the robot had. In particular, a highly transparent robot provided audible feedback about its status. However, they do not focus on how people perceive the performance of the robot with different levels of transparency. Boyce et al. [1] implemented an external interface (screen display) to make the operation of the robot more transparent. Their results showed that increasing transparency can help users understand a robot's environmental conditions and status. In contrast to both of these studies, we do not enhance the existing structure of robots. Instead we utilise their current setup as a way to keep the design of the robots as simple as possible.

Perception of motion in robots and interactive artefacts

Prior studies in HCI, HRI and UbiComp have examined whether people can infer intentionality, emotions or be motivated to interact with robots or artefacts through the visualisation of motion [15, 17, 10, 9, 2, 3]. Instead, in our study, we use the motion of a robot as a visual cue to change how people perceive the robot's performance. Closer to our work, Hoffman et al. [6] conducted a study where an anthropomorphic robot, *Travis*, was used as a speaker dock and music listening companion. Participants observed, listened and evaluated songs played by Travis. For some participants, Travis moved on-beat with the songs played. In contrast, other participants interacted with a moving Travis that was off-beat with the songs. The rest of the participants were introduced to a static Travis. Their results showed that participants rated songs significantly higher when the robot is moving on-beat with the songs than when it is static. Indeed, they pointed out the role of "personal robots as contributors to, and possibly amplifiers of, people's own evaluation of external events" [6].

These findings focus on the evaluation of events that are external to the system e.g. asking people whether they enjoy what they hear, instead of asking them about the quality of the sound produced by the system. Moreover, this work is about entertainment applications, while we look more at mundane or practical applications. In particular, we focus on how people evaluate the performance of such systems. Moreover, they centered their research on anthropomorphic robots. Instead, we are particularly interested with everyday systems (e.g. systems that are used in everyday situations such as cleaning or cooking robots). This is because it may not be practical to modify everyday systems to be anthropomorphic. We intend to focus on maintaining the simplicity of such systems.

Perception of robots motion through video and animation

Previous studies showed how people perceive robots through their physical movement [6]. However, there are other alternatives to interact with robots, such as, videos and animations. Such modalities allow people to visualise robots remotely without having a physical interaction with the system. Takayama et al. [21] examined people's perception of virtual animated robots through a lab study. For the study, the robot covered a variety of activities, such as opening a room, delivering a drink, requesting help from a person to plug into an outlet, and ushering a person into a room. Their results suggest that people are positively influenced by animations showing the outcome of a robot and more specifically that they read robot behaviour with more certainty. However, while the focus of their study is on training we are interested in real time interaction with robots. Additionally while their work is based on virtual animated robots, ours use physical ones. Wainer et al. [24] ran a study with participants that interacted in a collaborative task with an embodiment robot v.s. non-embodiment robot (e.g. simulated and video). In more detail, participants resolve a Towers of Hanoi puzzle following the instructions of the robots. Their results suggest that people perceive an embodiment robot more helpful and enjoyable in comparison with a non-embodiment robot. However, they did not analyse whether people perceive that one type of robot works better than the other, which we present as our key contribution.

MOTIVATION

A large body of work from cognitive psychology investigated how motion and other sensory cues influence our perception of the world. We started from this work to design notification mechanisms that could influence people's perception of autonomous robots.

In psychology "perception" is defined as the process that people follow to identify, interpret, and understand their environment, with the support of sensory (i.e. physical) and cognitive cues (referred to as *high-level of knowledge*) that the nervous system processes [18]. Studies have shown that humans can extract high level information from very basic motion cues [8]. However, in some cases, physical cues are insufficient for the brain to interpret the environment. Hence, the brain uses existing knowledge as a way to make sense of sensory signals (e.g. sight) [5].

Our perception of the world is sometimes influenced by more than one sensory channel. For example, McGurk and MacDonald demonstrated that speech perception is influenced by

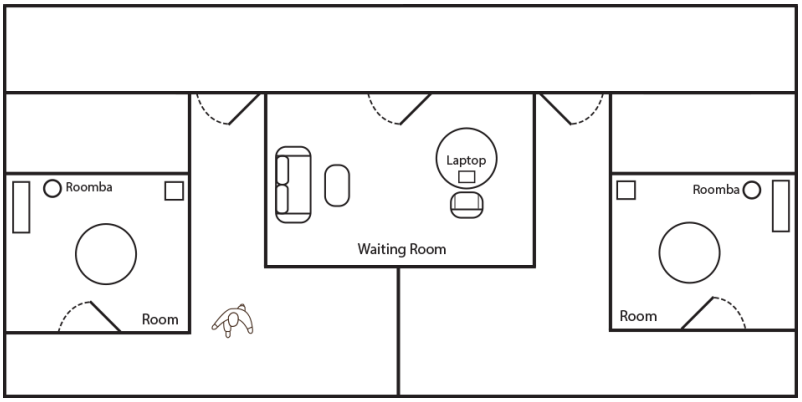

Figure 1. This figure shows a layout of the rooms where we conducted the experiment.

both sound and vision [14]. Vines et al. [23] reported a study where participants rated how much they liked audiovisual clips of clarinet players to investigate how different visual cues affect people's evaluation of the musicians' performance. They found that participants gave a lower score to a clarinettist who did not move compared to a clarinettist with more expressive body motion. This result suggests that an appropriate visual cue can improve people's rating of a non-visual property. Building on such a corpus, we set out to explore whether motion can be leveraged to influence people's perception of autonomous robots.

STUDY DESIGN

A user study was designed and conducted to analyse the effect of motion, as a visual cue, on people's perception. Specifically, the study was designed to test the following hypothesis:

> H1 – The *visualisation of automation*, which is showing system process through motion (that is intrinsic to the system), is an effective visual cue that can positively change people's perception of the system.

Experiment 1

Experiment Design

For this experiment, we selected the Roomba robot because it is an off-the-shelf product designed for domestic everyday use. A within-groups design was used to compare the effect of the visual cue and its absence on the same set of participants, where participants evaluated and compared the performance of two Roomba robots in vacuuming the carpets in two rooms where each was located. Two conditions were defined in our experiment: *no-motion* and *motion*. In the no-motion condition (the control condition), participants saw the robot after it completed its task, having already returned to its charging base. In contrast, in the motion condition (the 'treatment' condition) participants saw the robot moving as it docked in its charging base, having completed its cleaning duties. This movement is the *visual cue* at the centre of our study. In practical terms, the motion condition was implemented through a Wizard of Oz approach whereby an experimenter activated the robot seconds before participants arrived. The study was fully counterbalanced: half of the participants saw first the robot in the motion condition, vice versa for the other half. Moreover, the rooms and the robots were also alternated and fully counterbalanced: half of the participants saw the robot in the motion condition in Room A and the other half saw the robot for the same condition in Room B. The first robot participants saw was referred to as simply 'robot A', and the other one 'robot B', regardless of the condition (so that the naming would not influence the results).

At the beginning of each experiment, participants were told that the task was to compare two different algorithms implemented on each of the two robots. After this introduction, participants were asked to visit two rooms and were given a questionnaire asking them to evaluate on a 5-point likert scale the *cleanliness of the carpet* (from "1 - dirty" to "5 - clean without any chance to improve"). They were then asked to move to a different room to wait until the Roombas finished vacuuming the carpets. Participants were explained that they had to wait in a different room because the algorithms were still work-in-progress so we did not want their judgement to be influenced by their trajectories. Figure 1 shows the layout of the room. While waiting, participants were asked to play a puzzle game [2] on a 13" screen laptop. When the two robots had completed their tasks, participants received a text-based notification shown on the laptop indicating that they could go and evaluate the performance of the Roombas. After receiving the notification, participants visited both rooms one after the other. As described above, in one room they found the robot

[2] http://tinyurl.com/krx3w73

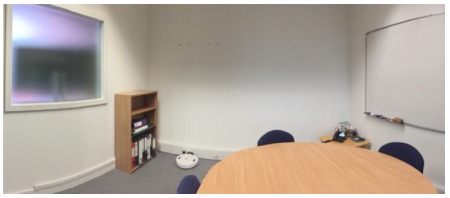

Figure 2. On this figure, we can see one of the rooms where the experiment took place.

already docked, while in the other they saw it docking. After participants had seen the robot docking, we told them that the robot's action of docking was not related to the robot's task of cleaning the carpet. As such, they were allowed to see this part of the robot's process.

After visiting each room, they were asked to continue the questionnaire and evaluate whether there is an *improvement with the cleanliness of the carpet* on a 5-point likert scale (from "stayed the same, did not have an improvement" to "better than before"). The post-task question was phrased differently from the pre-task questions, so the answers cannot be directly compared.

Once they evaluated both rooms, participants were asked to compare the performance of the Roombas. To try and ensure that participants would provide a significant and thoughtful evaluation, we designed a performance-based reward mechanism. Participants were asked which of the Roombas they thought most people would select as the one with the best performance (including the option that both had the same level of performance), and they were told that only if they selected the most popular choice at the end of the study (after we collected data from all participants) they would be rewarded with a £10 voucher (hereafter referred to as *reward-based question*). To check whether participants subjective judgement of the Roombas differed from what they expected the majority of people to choose, after they answered the first question they were presented with a second question, asking them which robot they personally consider to be the one with the best performance, regardless of other people's opinion. This second question (referred later as *non reward-based question*) had no effect on the reward received by the participants.

External validity was a key factor in the design of the experiment to test the effectiveness of visual cues in people's perception when they evaluate the performance of the robots. Therefore, we were particularly careful in keeping a number of variables that could affect participants' perception of the performance of the Roombas constant. These variables were determined through pilot studies:

- Cleanliness of carpets: The Roombas did not actually clean the carpets during the experiment.
- Robot's environment: The rooms used in the experiment were similar to maintain the same conditions (see Figure 2). Moreover, the robots were switched between the two rooms to maintain a fully counterbalanced study design.
- Roombas' task completion time: Both of the Roombas were simulated to vacuum the rooms in 10 minutes and were working simultaneously.
- Robot's model: The two robots used in the experiments were of the model iRobot Roomba 500.
- Evaluation time: Participants were only allowed 15 seconds to evaluate the carpets in each room. This was done to avoid participants spending more time in one room than the other.

Participants

A total of 16 participants (12 female, 4 male) took part in the study and 15 of these were members of the university: PhD and Masters students, none of which had technical background (e.g., not from Computer Science or Engineering). One participant was a homemaker. The ages of these participants ranged from 24 to 53 years old (M = 32.00, SD = 7.46).

Results

Selection of robot with best performance. For the *reward-based* question, 15 out of the 16 participants selected the moving robot (motion condition) as the one with the best performance. The remaining participant selected the robot in the no-motion condition as the best performing one, while nobody indicated that the robots had the same level of performance. For the *non reward-based* question, only one participant expressed a different opinion from that of the previous question, saying that both robots had the same performance. In total, 14 participants considered the moving robot as the better performing robot when answering the non reward-based question. These results are illustrated in Figure 4.

Cleanliness of the carpets. A Mann-Withney test revealed a statistically significant effect ($U = 67.50, p < .05, r = .41$) of the motion on the rating of how clean the rooms were after the operations of the robots. The room in the motion condition was rated on average as cleaner ($mdn = 2.5$) than the room in the no-motion condition ($mdn = 1.5$). Figure 3 shows the means comparison of the two groups. No statistically

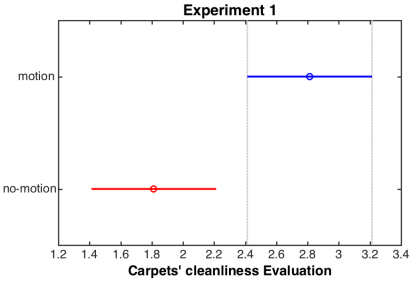

Figure 3. Comparison of evaluation means for rooms' carpet after robots clean, with 95% confidence confidence bars.

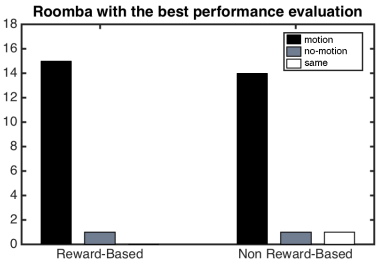

Figure 4. Comparison between participant's evaluation in the reward-based and non reward-based questions

significant differences were found on the ratings of how clean the rooms were before the operations of the robots between the conditions.

Discussion

The results of Experiment 1 confirms H1: motion can be used to change people's perception of how well an automated or automatic system works. The data shows clearly that the change is in the positive direction: all except 2 participants agreed that the robot in the motion condition was the one with the best performance. All except 1 declared that they thought the moving robot would be considered by other participants as the one with the best performance. This finding is further confirmed by the ratings that participants expressed through the likert scales. In the motion condition the room was rated as cleaner than in the no-motion condition, after the robots' operations. As expected, no differences were found on the rating of the rooms before the robots' operations. These results show that the visualisation of automation can be used as a tool to change people's perception of the performance of automated or automatic systems, even in the case of systems that are not anthropomorphic, extending what was previously reported in the literature [6].

Experiment 2

The results from the first experiment clearly show that seeing a robot moving had an effect on our participants' perception of its performance. However, it as far as I know noted that the movement was seen *in person*. Could the same effect be observed if the movement of the robot is experienced through a video feed? Indeed there might be situations in which users are unable to directly see the movement of a robot. To answer this question we designed a second experiment to compare how people perceive the performance of a Roomba when people watch a video of it docking in comparison to watching a Roomba docking in person. As such, we defined a new hypothesis:

> H2 – *Physical visual cues* are more effective than *video-based* cues at positively influencing how people evaluate an autonomous robot that show such cues.

Experiment Design

The design of experiment 2 is the same as experiment 1, except that the *no-motion* condition was replaced by a *video* condition. When participants received the notification that the robot had completed its task in the video condition, they were presented a video showing the Roomba docking. This is similar to the motion condition, but mediated over a video, rather physically seen in the same environment. In this new condition, a video of a Roomba docking was displayed on the laptop computer where the participants played the video game (cfr Experiment 1), and this served as a notification that the Roomba has completed its operation, rather than the text-based notification. For practicality the video was a pre-recorded clip, but it was presented to participants as a live feed from the room (the two rooms had no external windows, making such mockup realistic). Moreover, to avoid details on the video that can change people's perception we used a VGA resolution. To guarantee that participants associate the video-based notification with the correct Roomba the notification was presented to the participants before they visited the room. To accomplish this, the research investigator carried the laptop throughout the duration of the study, including when the rooms are about to be evaluated. Before entering the rooms where the Roombas were, the investigator would show the laptop's screen. For the video condition, this means that they would see the video-based notification right before they enter the corresponding room, therefore guaranteeing that they associate the video with the correct Roomba.

We included some new questions in the final questionnaire. In addition to the *reward-based* and *non-reward-based* questions, participants were also asked why the robot they selected performed better than the other, with a view to understand the motivation behind their choices. Moreover, they were asked whether they would prefer watching a video of the Roomba working or watching the Roomba physically finishing its task and why.

As in experiment 1, we were particularly careful in keeping a number of variables that could affect participants' perception of the system performance constant. These variables were the same as those listed for experiment 1.

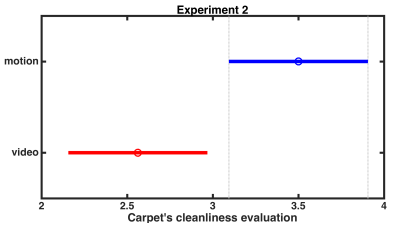

Figure 5. Comparison of evaluation means for rooms' carpet after robots clean, with 95% confidence confidence bars.

Participants

A total of 16 participants (10 male, 6 female) took part in the study and all of them were members of the university: undergraduate and postgraduate students, including a wide range, from Computer Science, English literature, Mechanics, Economics and Psychology students. The ages of the participants ranged from 19 to 37 years old (M = 22.00, SD = 4.39).

Results

Selection of Roomba with the best performance. In total 13 of the 16 participants considered that the Roomba in the motion condition performed better than the Roomba in the video condition. The remaining three participants indicated that the Roomba in the video condition performed the best, while nobody suggested that both robots had the same performance. All participants answered in the same way the *reward-based* and *non reward-based* questions, i.e. they all believed their answer would be the most popular one. These results are illustrated in Figure 7.

Reasons for choosing one Roomba over the other. The responses to the question about why participants selected a particular Roomba as the one performing the best were summarized through open coding. Each response was associated to one or two codes, with five codes used in total: *details, relative, generic, room features,* and *clean already*. Figure 6 illustrates the frequencies of these codes for those who preferred the motion condition and those who preferred the video condition. The code *details* was associated to responses which referred to specific issues in the room, such as "crumbs which lie close to chair legs" and "coffee stains." The code *relative* was used when the responses referred to the comparison of how clean the room was before and after the operation of the robots, such as: "Found the room cleaned by Roomba A much cleaner than it was initially" and "biggest change in cleanness". Comments coded as *generic* included "cleaned the room better" and "The carpet of room B was cleaner than room A". The code *room features* was used when participants referred to the influence of room features on the performance of the robots, such as "less corners for roomba to have difficulty with" and "It seemed to clean tighter spaces better". Finally, one participant stated that the room was clean to start with ("Because the Room B is clean already so it is hard to evaluate the Roomba B performance") so this response was coded as *clean already*.

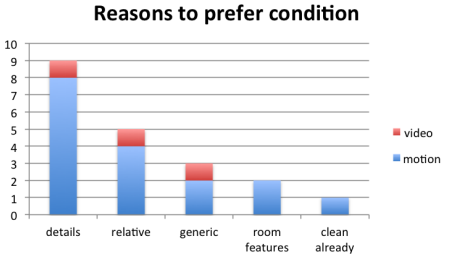

Figure 6. Reasons expressed by participants for preferring one Roomba over the other in Experiment 2.

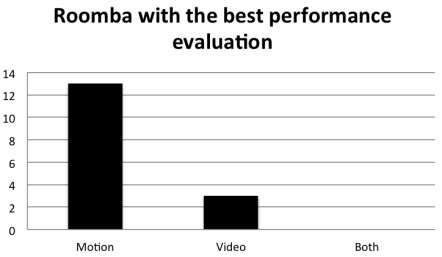

Figure 7. Comparison between participant's evaluation of the two different modalities. Note that all participants responded to the reward and non-reward based questions in the same way.

Cleanliness of the carpets. A Mann-Withney test revealed a statistically significant effect ($U = 70, p < .05, r = .39$) of motion and video on the rating of how clean the rooms were after the robots' operations. The room was rated on average as cleaner in the the motion condition ($mdn = 3$) than in the video condition ($mdn = 2$). Figure 5 shows the means comparison of the two groups.

Modality preference. Ten participants preferred the video over seeing the robot physically move; four participants preferred seeing the robot in person; while the remaining two participants did not have a preference for how they see the robot.

Reason for preferring a modality. The responses to the question about why participants selected one modality over the other were summarized through open coding. Each response was associated to one code, with six codes used in total: *better understanding, convenience, emotional, generic, reliable* and *subjective*. Figure 8 illustrates the frequencies of these codes. An example in the *better understanding* category included "I can understand which part of the room have been cleaned". The *convenience* category included "I do not have to be there till the end", "Can observe the room situation remotely", and "This will save our time while we are doing some other work during the time Roomba was doing its task...". Comments categorised as *emotional* included "fun" and "...physical presence has a more personal effect". Comments in the *generic* category included "You can see the Roomba working physically and the video is helpful" and "Able to see functionality of the roombas". An example comment in category *reliable* category included "On the video you can't see what is happening". Finally, the *subjective* category included "I'm personally a visual person so it illustrates it much better...".

Discussion

The results of Experiment 2 suggest that seeing the robot moving in person positively influences the perception of its performance, compared to seeing it through video. All except

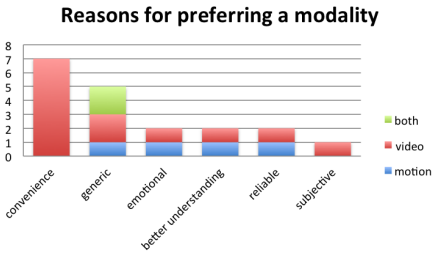

Figure 8. Reasons expressed by participants for preferring one notification modality over the other in Experiment 2.

3 participants agreed that the robot in the motion condition was the one with the best performance. This finding confirms our hypothesis H2. Our participants' ratings of the cleanliness of the rooms further confirm such result: in the motion condition the room was rated as cleaner than in the video condition, after the robots' operations.

The qualitative data about why participants selected one specific robot as the one performing better provide further evidence of the effect of the two different notifications and their potential to influence people's perception of the robots. As illustrated in Figure 6, participants provided generic answers to this question only in three instances. In contrast, in the majority of cases our participants' answers included specific and tangible reasons to support their choice, despite the fact that the Roombas did not actually clean either of the two rooms.

Even though the performance ratings clearly indicate that the Roomba in the motion condition is the most successful one, when participants were asked about their general preference regarding the modality most of them chose the video. The most frequent reason to support this choice was convenience. Such contrast between performance ratings and general preference seems to suggest that participants were not aware of the bias that the motion condition caused on their performance rating. It should also be noted that while the performance rating was related to a financial incentive the question about general preference was not. Therefore it is also possible that participants answered the latter more casually.

GENERAL DISCUSSION AND IMPLICATIONS

The results of both of our experiments indicate that the feedback delivered with notifications can have a considerable influence on people's perception of the performance of autonomous robotic systems. In particular, seeing the robot moving as it finishes its operation in person led our participants to rate its performance higher than not seeing any motion, or seeing the same motion over video.

Compared to prior work [6, 22], the type of motion displayed in our study is very simple, making it very easy and cost effective to take advantage of our findings in existing designs. Indeed, in the case of the Roomba, the motion cue we studied is simply part of the standard operation of the robot. However, additional measures may need to be put in place to drive the user's attention to the motion. For example, presence or location sensing (including e.g. smartphone apps to detect the user's location) may be employed to activate the system when users are physically close to them, or on their trajectory home, leveraging prior work on pattern recognition on GPS traces [7].

These results could potentially apply to a wide range of devices. In the domestic context, smart appliances such as vacuum cleaning robots, washing machines (e.g., seeing the spin cycle confirms the clothes will be clean and dry) and dishwashers (e.g., hearing the dishwasher rinse and shut down confirms all dishes have been cleaned) could be timed according to GPS traces such that when they detect (or predict) that the owners are nearby, they would finish their cycles [7]. A similar approach could also be used for prototyping machines, such as 3D printers, laser cutters and CNC machines.

In addition, the results of our experiments highlight new research opportunities around different ways to present visual cues as new forms of feedback for autonomous robots. While our results, even though on a small sample, show a clear effect, they also open a number of new research questions, for example: Is this effect long lasting? Does it apply to any kind of robots, or even other ubicomp (non-robotic) systems? Is the timing of the cues that are presented important? We believe that the effect we observed may even influence people's inclination to adopt such systems: more research is required in such a direction.

CONCLUSION

In this paper, we have presented two lab experiments, each with 16 participants, designed to investigate whether seeing the motion of autonomous robots in person can positively change people's perception about the performance of the robots. Indeed, our findings suggest that people's perception of the performance of an autonomous robot can be improved for the better through showing them the robot moving, in such a way that they would see it in person. Showing the motion of autonomous systems acts as a visual cue to help people perceive the performance of such systems correctly. In contrast to previous work, our results apply to systems which are not anthropomorphic, hence, the implications can be relevant to a large number of systems. Therefore, we hope that the results presented in this paper will stimulate designers to integrate motion in the feedback of their systems, and researchers to further explore this area.

ACKNOWLEDGMENTS

This work was supported in part by CONACyT, SICYT Morelos, and EPSRC ORCHID and A-IoT, and approved by University of Southampton Ethics Committee (ref: 17155). The data referred to in this paper can be found at http://dx.doi.org/10.5258/SOTON/397976

REFERENCES

1. Michael W. Boyce, Jessie Y.C. Chen, Anthony R. Selkowitz, and Shan G. Lakhmani. 2015. Effects of

Agent Transparency on Operator Trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts (HRI'15 Extended Abstracts)*. ACM, New York, NY, USA, 179–180. DOI: http://dx.doi.org/10.1145/2701973.2702059

2. Mason Bretan, Guy Hoffman, and Gil Weinberg. 2015. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human-Computer Studies* 78 (2015), 1 – 16. DOI: http://dx.doi.org/10.1016/j.ijhcs.2015.01.006
3. Anca D. Dragan, Shira Bauman, Jodi Forlizzi, and Siddhartha S. Srinivasa. 2015. Effects of Robot Motion on Human-Robot Collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*. ACM, New York, NY, USA, 51–58. DOI: http://dx.doi.org/10.1145/2696454.2696473
4. Doris A Fuchs. 2003. *An institutional basis for environmental stewardship*. Vol. 35. Springer Science & Business Media.
5. V Gregory. 1998. Eye amd Brain, the Psychology of Seeing. (1998).
6. Guy Hoffman and Keinan Vanunu. 2013. Effects of Robotic Companionship on Music Enjoyment and Agent Perception. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction (HRI '13)*. IEEE Press, Piscataway, NJ, USA, 317–324. http://dl.acm.org/citation.cfm?id=2447556.2447674
7. Eric Horvitz and John Krumm. 2012. Some Help on the Way: Opportunistic Routing Under Uncertainty. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. ACM, New York, NY, USA, 371–380. DOI: http://dx.doi.org/10.1145/2370216.2370273
8. Gunnar Johansson. 1973. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics* 14, 2 (1973), 201–211. DOI: http://dx.doi.org/10.3758/BF03212378
9. Jinyung Jung, Seok-Hyung Bae, and Myung-Suk Kim. 2013a. Three Case Studies of with Moving Products. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '13)*. ACM, New York, NY, USA, 509–518. DOI:http://dx.doi.org/10.1145/2493432.2493442
10. Jinyung Jung, Seok-Hyung Bae, Joon Hyub Lee, and Myung-Suk Kim. 2013b. Make It Move: A Movement Design Method of Simple Standing Products Based on Systematic Mapping of Torso Movements & Product Messages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1279–1288. DOI: http://dx.doi.org/10.1145/2470654.2466168
11. Frédéric Kaplan. 2005. Everyday Robotics: Robots As Everyday Objects. In *Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies (sOc-EUSAI '05)*. ACM, New York, NY, USA, 59–64. DOI: http://dx.doi.org/10.1145/1107548.1107570
12. Jun Kato, Daisuke Sakamoto, Takeo Igarashi, and Masataka Goto. 2014. Sharedo: To-do List Interface for Human-agent Task Sharing. In *Proceedings of the Second International Conference on Human-agent Interaction (HAI '14)*. ACM, New York, NY, USA, 345–351. DOI: http://dx.doi.org/10.1145/2658861.2658894
13. Taemie Kim and P. Hinds. 2006. Who Should I Blame? Effects of Autonomy and Transparency on Attributions in Human-Robot Interaction. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*. 80–85. DOI: http://dx.doi.org/10.1109/ROMAN.2006.314398
14. Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (23 12 1976), 746–748. http://dx.doi.org/10.1038/264746a0
15. Ditte Hvas Mortensen, Sam Hepworth, Kirstine Berg, and Marianne Graves Petersen. 2012. "It's in Love with You": Communicating Status and Preference with Simple Product Movements. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems (CHI EA '12)*. ACM, New York, NY, USA, 61–70. DOI: http://dx.doi.org/10.1145/2212776.2212784
16. Donald A Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.
17. Diana Nowacka, Nils Y. Hammerla, Chris Elsden, Thomas Plötz, and David Kirk. 2015. Diri - the Actuated Helium Balloon: A Study of Autonomous Behaviour in Interfaces. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 349–360. DOI:http://dx.doi.org/10.1145/2750858.2805825
18. Daniel L. Schacter, James Eric Eich, and Endel Tulving. 1978. Richard Semon's theory of memory. *Journal of Verbal Learning and Verbal Behavior* 17, 6 (1978), 721 – 743. DOI: http://dx.doi.org/10.1016/S0022-5371(78)90443-7
19. D. Schulz, W. Burgard, D. Fox, S. Thrun, and A.B. Cremers. 2000. Web interfaces for mobile robots in public places. *Robotics Automation Magazine, IEEE* 7, 1 (Mar 2000), 48–56. DOI: http://dx.doi.org/10.1109/100.833575
20. J. Y. Sung, R. E. Grinter, H. I. Christensen, and L. Guo. 2008. Housewives or technophiles?: Understanding domestic robot owners. In *Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on*. 129–136.

21. Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing Thought: Improving Robot Readability with Animation Principles. In *Proceedings of the 6th International Conference on Human-robot Interaction (HRI '11)*. ACM, New York, NY, USA, 69–76. DOI: http://dx.doi.org/10.1145/1957656.1957674

22. Jo Vermeulen, Kris Luyten, and Karin Coninx. 2013. Intelligibility Required: How to Make Us Look Smart Again. (2013). DOI:http://dx.doi.org/1942/15998

23. Bradley W. Vines, Carol L. Krumhansl, Marcelo M. Wanderley, and Daniel J. Levitin. 2006. Cross-modal interactions in the perception of musical performance. *Cognition* 101, 1 (2006), 80 – 113. DOI: http://dx.doi.org/10.1016/j.cognition.2005.09.003

24. J. Wainer, D. J. Feil-Seifer, D. A. Shell, and M. J. Mataric. 2007. Embodiment and Human-Robot Interaction: A Task-Based Perspective. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*. 872–877. DOI:http://dx.doi.org/10.1109/ROMAN.2007.4415207

25. 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

Appendix B

HCI Journal Paper

In this appendix, we present the submitted version of the HCI journal paper.

ARTICLE TEMPLATE

Seeing (Movement) is Believing: the Effect of Motion on Perception of Automatic Systems Performance

A. N. Author[a] and John Smith[b]

[a]Taylor & Francis, 4 Park Square, Milton Park, Abingdon, UK; [b]Institut für Informatik, Albert-Ludwigs-Universität, Freiburg, Germany

ARTICLE HISTORY
Compiled May 26, 2017

ABSTRACT
In this paper, we report on one lab study and seven follow-up studies on the crowdsourcing platform designed to investigate the potential of animation cues to influence users perception of two smart systems: a handwriting recognition and a part-of-speech tagging system. Results from the first three studies indicate that indeed animation cues can influence a participants perception of both systems performance. The subsequent three studies, designed to try and identify an explanation for this effect, suggest that it is related to the participants' mental model of the smart system. The last two studies were designed to characterise the effect more in detail, and they revealed that different amounts of animation do not seem to create substantial differences and that the effect persists even when the systems performance decreases, but only when the difference in performance level between the systems being compared is small.

KEYWORDS
Sections; lists; figures; tables; mathematics; fonts; references; appendices

1. Introduction

There is a growing number of *smart systems* that help users to gather data and process information from sensors. These are systems that utilize some form of pattern recognition, machine learning, or more generally artificial intelligence to complete a variety of information-processing tasks. Until recent times, such smart systems were only accessible at high-cost for specialised applications (e.g. in medical fields, aviation), but more recently they have become increasingly widespread for non-specialist applications, such as apps that help people with office work (e.g. translation platforms[1] or document scanning[2]). As smart systems become available to a wider variety of users, it is important to study how *casual users* interact with them. Given that smart systems can involve advanced concepts in pattern recognition (e.g. Bayesian classification (Talbot, Lee, Kapoor, & Tan (2009))), or even act as black boxes (Krause, Perer, & Ng (2016)), their operation may be difficult to grasp for non-specialist users, who do not receive training (as it is common for domestic appliances).

CONTACT A. N. Author. Email: latex.helpdesk@tandf.co.uk
[1]`https://www.apertium.org`
[2]`https://www.camscanner.com/`

Research in psychology and behavioural economics indicates that people's perception and decisions can be influenced by cognitive biases, implemented often through nuanced cues. As such, we are interested in whether cues, and in particular *visual animation cues*, can influence users perception of smart systems. In particular, we focus on whether these cues affect how people rate the performance of a smart system. Indeed, recent research (Garcia, Costanza, Ramchurn, & Verame (2016)) demonstrated that simple *motion cues* can have quite a radical impact on people's perception of vacuum cleaning robots: when the interaction was orchestrated in such a way that participants saw the robot moving, they perceived it to clean better than a robot which worked identically, but was not seen moving. Building on this prior work, our aim is to investigate whether a similar effect can be noticed for GUI-based smart systems, through the use of *animation cues* integrated in the system interface. We argue that being aware of and understanding such biases are important for the design of interaction around smart systems. On one hand, it may reveal opportunities to improve users' perception e.g. making the system more popular or more likeable. On the other hand, and perhaps even more importantly, being aware of such biases may allow designers to avoid unintentionally deceiving users.

In this paper, we report on one lab study and seven follow-up studies on the crowdsourcing platform Amazon Mechanical Turk[3] (MTurk) designed to investigate the potential of animation cues to influence users' perception of two smart systems: a handwriting recognition (HWR) and a part-of-speech (POS) tagging system. All studies collected both quantitative and qualitative data, and used *consensus-oriented financial incentives* to increase ecological validity and motivate participants to provide thoughtful responses. Results from the first three studies indicate that indeed animation cues can influence a participant's perception of both systems' performance. Both in the lab and on MTurk, participants reported that the system, which had animation integrated in its UI, performed better than an alternative system, which was in fact identical apart from the animation. The subsequent three studies were designed to try and identify an explanation for this effect; their results suggest that the effect of animation cues is related to a participants mental model of the smart system. More precisely, if the cues are compatible with users' mental model of how the system works, they seem to somehow provide a reassurance about the system operation, and evoke an illusion that the system works better than an alternative one for which animation is not shown. Having identified a possible explanation for this phenomenon, we report two further studies designed to characterize it more in detail. In particular, these last two studies revealed that different amounts of animation do not seem to create substantial differences, and that the effect persists even when the system's performance actually decreases, but only when the difference in performance level between the systems being compared is small. These results support specific *implications* for the design of user interfaces for smart systems, which are discussed following the detailed report of the experiments.

2. Related Work

Our research aims to analyse how simple animations can change participants' perception of smart systems' performance. To this end, in what follows, we survey prior research that has studied cognitive biases and how different framings of information

[3] `https://www.mturk.com`

impact people's perception. Then, we discuss transparency for the intelligibility of software systems and how their design influences how people perceive the system. Finally, we discuss prior work on the perception of motion in screen-based systems.

2.1. *Cognitive Biases*

Studies in psychology and behavioural economics have shown that people's perception of how well a system or process works can be influenced by different cognitive biases. For example Tversky & Kahneman (1985) showed that people can be influenced by the way outcomes are described to them. In a survey, participants were presented with a problem and two possible solutions. These two solutions had the same outcome, however, one emphasised its positive aspects, while the other emphasised the negative aspects. Results suggest that people had a tendency to choose the solution that emphasised the positive aspects. As another example, Ariely (2008) ran a study to analyse if the price on medicine has a placebo effect on people's perception of how they feel after they took medication. One group received the medicine with the actual price and a second group received the medicine with a 10 cents discount (off an original price of $2.50). The results showed that while almost all participants in the first group experienced pain relief from the pill, *only half* of the participants who were given the "discounted medicine" experienced pain relief. In our work, we are interested in exploring whether there are also cognitive biases that can influence people's perception of how well smart systems work.

2.2. *Transparency and Intelligibility of Software Systems*

Prior research has examined the effect of increased intelligibility on people's understanding of smart systems. In particular, previous studies have suggested that smart systems should generate and provide meaningful explanations for systems' actions, behaviour or outcomes (Lim, Dey, & Avrahami (2009); Lyons (2013); Tullio, Dey, Chalecki, & Fogarty (2007)). For example, Lim et al. (2009) ran two experiments to analyse the effect of meaningful explanations describing *why* and *why not* a context-aware application behaved in a certain way. Their findings suggest that users have a better understanding of a system's behaviour and a higher feeling of trust in it when it provides explanations. Moreover, Tullio et al. (2007) ran a six-week field study to analyse whether intelligibility can help office workers improve their understanding of how a system predicts their managers' interruptibility. They found that people were able to understand the system prediction better, even if the overarching structure of their mental model stayed stable during the study. However, explanations can also cause information overload, possibly confusing and overwhelming users (Lim & Dey (2011); Yang & Newman (2013)). Similarly, another study investigated *Laksa*, a context-aware software which used eight question type explanations (e.g. *Why, Why Not, What If*) to explain its decision to the users (Lim & Dey (2011)). To evaluate the software, participants used the software in three situational dimensions (exploration, fault finding, and social awareness) that allowed the researchers to observe whether participants do or do not understand software decisions. They noted that quickly consumable explanations of a systems output are crucial and additional, richer explanations should be easily accessible. Lim & Dey (2011) observed that prior knowledge plays an important role in both understanding of such systems and also interpreting the explanations given. The lack of previous knowledge can lead people to misunderstand or misuse a

system. Complementing this prior work, our aim is to understand whether it is possible to change people's perception of smart systems without increasing their cognitive workload by, for example, providing additional cues (e.g. through animation) that can expose to users that a smart system is doing work.

Another way of improving system intelligibility is through information visualisation, which is the use of visual representations of data structures and algorithms to help people analyse data (Card, Mackinlay, & Shneiderman (1999); Ware (2012)). The concept of information visualisation is considered a method to make a system understandable without providing explanations of its process. For example, O'Donovan et al. (2008) ran a study where participants interacted with *PeerChooser*, an interactive visualisation system for collaborative filtering. The system generated a peer-graph which is centred on the current user. The graph showed a visual representation of their peer group or neighbourhood allowing participants to manipulate connections with their neighbours. This interaction allowed participants to visualise recommendations from the system based on their preferences. Their findings suggest that a visual-interactive approach can improve the accuracy of the recommendations provided by the system and also enhance user experience (O'Donovan, Smyth, Gretarsson, Bostandjiev, & Höllerer (2008)). In our case, instead of using interactive visualisations, we explore visualisations of a system's process through motion (animations) that represents its execution of a task.

An example of a study that uses motion as a visual feedback to explain a system's decision is presented by Vermeulen (2010). In their study, animations were used to show the process that a system follows when it makes a decision, given how a user interacts with its inputs (e.g., switch) or sensors (e.g. motion detector). Findings from their study suggest that participants understood the decisions and actions taken by the system because of the explanations they received. This approach demonstrates that animation, as a feedback, can help people understand decisions made by a system. However, participants also found it difficult to track the animation at times, thereby confusing them. We build on this idea and want to further explore how people's perception changes depending on the animation.

2.3. *The role of motion in users' perception of systems*

Research has looked at how people perceive motion in screen-based systems. *Animacy*, as Tremoulet & Feldman (2000) state, is when people perceive an object as being alive, through the pattern of its movements. They mention that the movement of an object does not need to be dramatic to show animacy (Fritz Heider (1944); Reeves & Nass (1996)). As a consequence, people attribute motivations, or intention in objects' movements from the patterns that these objects follow. This means that people can infer objects' intentions through their movements (Gao & Scholl (2011); Michotte (1963); Pantelis & Feldman (2012); Schlottmann & Surian (1999)). This has also been observed durng people's interaction with physically actuated interfaces such as helium balloons (Nowacka, Hammerla, Elsden, Plötz, & Kirk (2015)) or vacuum cleaning robots (Garcia et al. (2016)). Therefore, through designing the movement, it is possible to affect how people perceive objects. Michotte (1963) showed in their study that if two objects are in the same frame and suddenly change their direction, people can infer that both objects have a causal interaction. Pantelis & Feldman (2012) ran a study with multiple objects moving around on a screen. They found that, after watching multiple objects moving on a screen, people make interpretations of the intention and

behaviour of the objects. Moreover, in their experiment, people were able to distinguish if an object behaved friendly or hostile when it was moving around other objects. This body of work makes us believe that - by showing people an animation - they can be convinced that a system is working on a task. As such, we presume that people perceive a system that somehow communicates that it is doing work perform better than a system that hides how it works.

However, it has also been shown that some features of animations can confuse people and negatively impact people's perception. These features include but are not limited to: interaction between multiple objects (Gao, McCarthy, & Scholl (2010)), trajectories that are too complicated (Dittrich & Lea (1994); Tremoulet & Feldman (2000)), unnatural movements (Popović, Seitz, & Erdmann (2003)), or static backgrounds that are too complex (Gelman, Durgin, & Kaufman (1995)). Hence, it is important to ensure such issues are avoided when providing feedback about a system's execution of tasks.

Prior research has also analysed affective qualities of an interface depending on how the information and motion are presented on a screen (Detenber & Reeves (1996); Park & Lee (2010b)). Park & Lee (2010a) ran a study to understand how motion (i.e. transition effects between objects) influences the affective quality of an interface to improve user experience. They presented an image viewing interface that allows users to browse through a set of photos as they shift horizontally from one to another. Their results show that motion influenced how people rated affective qualities of the interface (e.g. youthfulness, calmness, and uniqueness). Also related to the effect of animation on user emotion, Bakhshi et al. (2016) reported that social network users have a tendency to share content more frequently if it involves animations, compared to content that is purely static. In contrast to this prior work, our interest lies in observing if motion has an effect on how people perceive systems' performance rather than on people's emotions.

3. Study 1

Building on the prior work discussed above, we set out to assess the potential for animation cues to influence users' perception of the performance of smart systems. We consider this investigation as an attempt to generalize the results by Garcia et al. (2016), who reported that motion could influence the perception of the performance of vacuum cleaning robots.

We designed a lab study involving a relatively simple graphic animation, somewhat related to the system operation. We chose a HWR system, a system that recognises handwritten text and converts it to electronic text (or e-text / typed text), because this is a common task that many people can relate to, at least conceptually, and it also can be simulated easily (Verame, Costanza, & Ramchurn (2016)). Moreover, we chose to use text in Filipino, a language that most users would be unlikely to know, to mimic the likely circumstances of casual users not being familiar with the kind of data handled by the system. In this way, rather simply checking the system output for typos, users are required to compare the input and the output looking for differences, a task that is more attention demanding.

3.1. *Method*

3.1.1. Study Design

A fully counterbalanced, within-participants design was used, where participants were asked to evaluate and compare the performance of two HWR systems, each corresponding to an experimental condition: *animation* and *no-animation*. Both systems were based on the same graphical user interface, illustrated in Figure 1. On the left-hand side of the screen, a scan of a page of handwritten text in Filipino (system's input) is displayed, while on the right-hand side the typed text (system's output) is shown. In both cases, the interface screen was preceded by a 'loading' screen, showing just a text that the system was processing its task for 10 seconds, to reinforce the idea that the system was doing something in the background. In the *no-animation* condition, the system presented the result immediately after the loading screen, and no motion cues were displayed. In the *animation* condition, after the loading screen, an animation was shown: on the last two lines of the input words were highlighted one by one, with a delay of a few hundred milliseconds; as each handwritten word was highlighted, the corresponding word on the output appeared. The first word highlighted by the animation was the word "naging", for more detail see Figure 1. By highlighting the words one by one, our intention was to give users an impression of how an algorithm may process the input data[4].

External validity was a key factor for this study. Therefore both HWR systems showed the same handwritten text, and both systems involved the same number of errors (four mistakes per paragraph, resulting in a total of eight mistakes across two paragraphs). In the last two sentences (the ones highlighted by the animation) both systems presented one error. Additionally, a consensus-oriented reward mechanism was adopted to try and ensure that participants would provide a meaningful and thoughtful evaluation when they choose which system they considered to have the best performance. Participants were told that if they select the system which the majority identified as the one with the best performance, they will be rewarded with a £10 voucher at the end of the experiment . This question is later referred to as *reward-based question*.

[4] https://vimeo.com/183480644

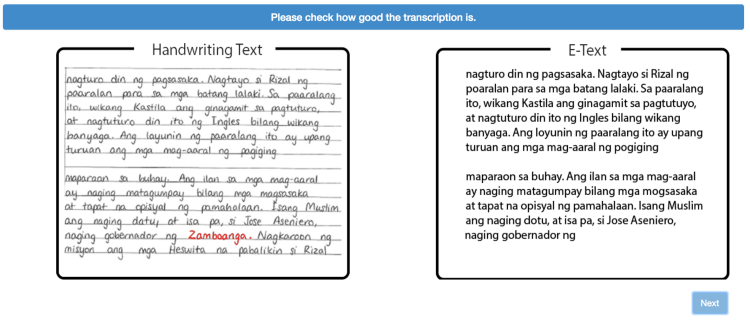

Figure 1. Interface of the HWR system implemented in Study 1.

3.1.2. Participants

A total of 16 participants (10 male, 6 female) took part in the study, and all of them were students: undergraduate and postgraduate, from a wide range of backgrounds, from Computer Science to Mechanics, Psychology and Design. Participants were recruited through adverts posted on university social network groups. The age of participants ranged from 18 to 24 years old (M= 20.68, SD=1.70).

3.1.3. Equipment

The study was run in a room at a university, where each participant sat with the investigator. The interfaces and the questionnaire were implemented as a simple Web application, using HTML5 and Python with the Django framework. The application was displayed on a 13" laptop, and served from the same computer each time. The animation was a GIF image.

3.1.4. Procedure

At the beginning of the study, participants received written instructions asking them to evaluate and compare the performance of the two HWR systems. The two systems were presented one at a time, in sequence: half of the participants first experienced the *animation* condition, while the other half first experienced the *no-animation* condition. In each condition the system was shown to participants for two minutes, so they had a limited time to compare input and output. After the participants had seen both systems, they were asked to fill in a questionnaire to evaluate their performance. Participants were firstly asked to rate the individual performance of each system on a 5-point Likert scale. They were then asked to select which of the systems they believe the majority of participants would choose to have the best performance and to provide a justification for their selection. As mentioned above, if participants answered in the majority they would receive a reward voucher. Additionally, we asked participants to describe in a text field why they chose one system over the other. After these questions were answered, participants were also asked (on a separate page) which system they considered to have the best performance without considering what the majority of the participants would choose, and no reward (we refer to this as the *non-reward-based question*).

3.2. *Results*

3.2.1. Selection of the system with the best performance

For the *reward-based* question, 12 of the 16 participants (75%) chose the system in the *animation* condition as the one with the best performance, the remaining 4 participants (25%) chose instead the one in the *no-animation* condition, while nobody indicated that both systems had the same performance level. Only 1 participant answered the *non-reward-based* question differently than the *reward-based* one, changing the answer from "no-animation" to "both systems". These results are illustrated in Figure 2.

3.2.2. Reasons for choosing one system over the other

Participants' responses to the question about why they selected a particular HWR system as the one that performed best were categorised through thematic analysis

Braun & Clarke (2006). Each response was associated to one or two of the following three themes: *number of errors, type of errors,* and *generic*. Figure 3 illustrates the frequencies of these themes. The theme *number of errors* was associated to responses where the participants reported finding fewer errors or mistakes in the output of one system than in the output of the other, such as "There were less mistakes in total", "It has mistakened less characters." and "I think both of them had about the same number of errors, however the second one's were more obvious [..]" Comments categorised as *type of error* were linked to situations when participants pointed out typographical errors they found, such as "only confuses a-o, b-h, ri-n whereas the second also confuses d-g" and "[..] algorithm only got mistakes when the words contain 'a' and 'o'." Finally, comments such as "More sensitive recognition of lettering [...]", and "[...] Errors of the second program are easier to guess and find out." were categorised as *generic*.

3.2.3. Performance ratings

A Wilcoxon Signed-rank Test revealed that the performance evaluation was higher for the *animation* condition ($Mdn = 4$) than for the *no-animation* condition ($Mdn = 3.5$), ($Z = 2.07, p < .05, r = 0.37$). Figure 4 shows participants evaluation of the performance of the systems.

3.3. *Discussion*

The results of Study 1 show that animation cues have an effect on participants' perception of the system's performance. The data shows clearly that the majority of participants considered the performance of the system in the *animation* condition to be better. It should be noted that this was the case despite the fact that one error was present in the sentence highlighted by the animation. In other words, even though the animation could have drawn the participants' attention to the mistake, for most of them the animation instead had the opposite effect. The qualitative data further supports this result, most participants seem to believe that the system in the *anima-*

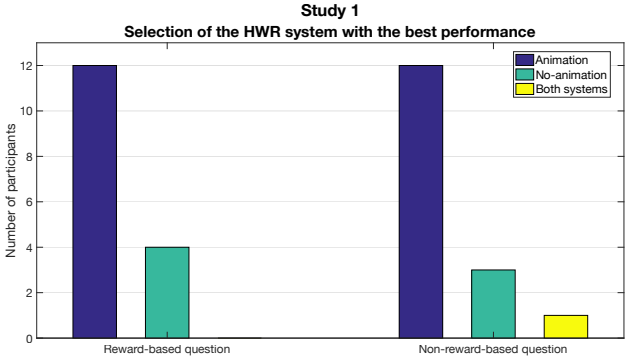

Figure 2. Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 1. Number of participants on the y-axis.

tion condition made fewer errors or different kind of errors than the other system, despite the two systems producing the same number and kind of errors. Moreover, participants seemed to be unconscious of the effect: none of the comments referred explicitly to the animation.

These results extend those from Garcia et al. (2016), who showed that motion can be used to change people's perception of the performance of vacuum cleaning robots. Our results indicate that the effect of motion does not apply only to physically moving systems, but also to graphical user interfaces through animation.

These results open up a number of follow up questions. Can this effect be observed in a less controlled environment? Can it be observed for a different type of smart system? The following two experiments were designed and carried out to address these

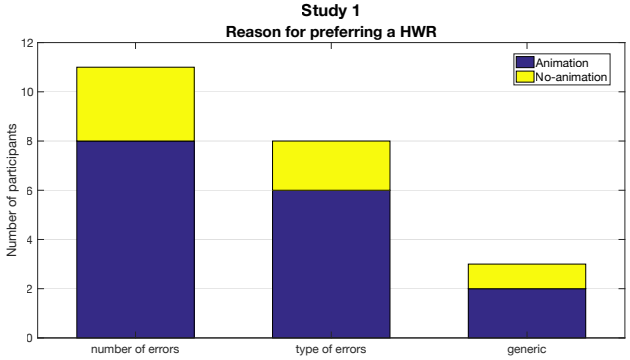

Figure 3. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition in Study 1. Number of participants on the y-axis.

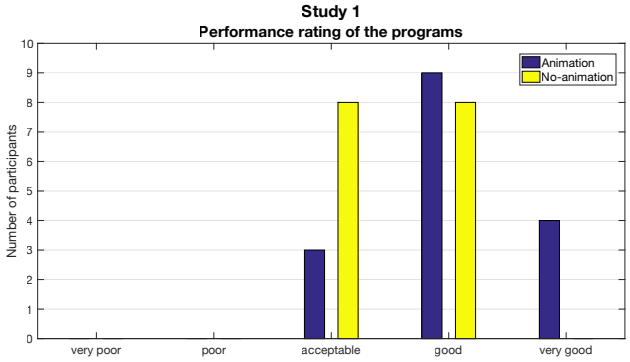

Figure 4. Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

two questions.

4. Study 2

To assess whether similar results to those of Study 1 can be observed also in a less controlled environment than the lab, we decided to run a similar experiment on a crowdsourcing platform: Amazon Mechanical Turk [5] (MTurk). In addition to being less controlled, crowdsourcing environments are also reported to include more diverse participants (Buhrmester, Kwang, & Gosling (2011); Germine et al. (2012)). Crowdsourcing has become a widespread online tool that researchers and companies use to outsource micro-tasks. In MTurk, these micro-tasks are referred to as Human Intelligence Tasks (HITs)[6], that leverage human computation, gather distributed and unbiased data, or validate results (Difallah, Catasta, Demartini, Ipeirotis, & Cudré-Mauroux (2015); Kazai, Kamps, & Milic-Frayling (2013); Mason & Watts (2010)). The people who complete such crowdsourcing tasks are referred to as 'crowd workers,' or simply 'workers.'

4.1. *Method*

4.1.1. Study Design

The study design was the same as in Study 1: fully counterbalanced, within-participants, where participants were asked to evaluate and compare the performance of two HWR systems, each corresponding to an experimental condition: *animation* and *no-animation*. The two conditions were identical to Study 1. However, we added an extra question in the post-task questionnaire asking participants to justify their selection for the non-reward-based question.

Because of the constraints of the MTurk platform, the reward mechanism was adjusted accordingly. Participants received a *fixed reward*, to compensate them for the time they spent working on our study, as well as an additional *performance-based, consensus-oriented reward*, designed to increase the ecological validity of the study and motivate the participants to provide thoughtful responses, similar to Study 1. To make sure that participants received a fair payment, we considered the minimum wage across the different countries participants could be from (see restrictions below), and we selected the Canadian one as the one currently highest, at approximately $10 per hour. Therefore, the fixed reward was set to $1.17 – the whole task takes about 7 minutes, $1.17 corresponds to about 10 minutes, leaving some margin. This amount was awarded as soon as all participants completed the study. The performance-based, consensus-oriented reward was awarded as a "bonus" on the MTurk platform, and it amounted to the same value as the fixed reward. In other words, the performance-based, consensus-oriented reward doubled the money that participants received for the study. It was awarded once all participants had completed the study.

4.1.2. Participants

Participants were recruited through MTurk, with two restrictions. First, they were only allowed to take part in the study if their location was *United States, Canada,*

[5] https://www.mturk.com

[6] `https://www.mturk.com/mturk/help?helpPage=overview`

or Australia, to avoid issues related to English comprehension. Second, recruitment was limited to participants with HIT approval rate was equal to 100% (this is the approval from those who advertise the HITs[7]), as rejection on MTurk often indicates that workers do not take tasks seriously. A sample of 16 participants successfully completed the online study, to keep the same pool size as Study 1 so that we were able to compare the two studies.

The age of participants ranged from 21 to 44 ($M = 33, SD = 5.94$), with 12 of them being males (75%) and 4 being females (25%). All our participants were United States nationals. The education levels of the participants ranged from secondary school level to master's degree or (equivalent). Overall 10 of them had a university degree level, 5 a secondary school level and 1 master's degree level. None of our participants reported knowing Filipino.

4.1.3. Equipment

We used the same Web application developed for Study 1. However, this was extended with an initial questionnaire to obtain the participants demographic information, and it was deployed to a publicly accessible Web server, to allow MTurk' workers to access it from their personal computers. On the MTurk platform, we added the URL link in the HIT where workers could access it.

4.1.4. Procedure

Before participants accepted the task, they were told that the aim of the experiment was to compare two different HWR systems, one at a time[8]. They were instructed to check the system's outcome and find possible mistakes that the system could make in the transcription of the handwritten text to typed text. After the introduction, the participants can decide to either accept or reject the task. Once they decided to accept the task, an external link was displayed. The link opened a new window that showed a brief questionnaire, asking for the participants' demographic information.

After the participants answered the initial questionnaire, the study followed the same procedure as Study 1, with the addition of an extra question in the post-task questionnaire (asking participants to justify their selection for the non-reward-based question), as mentioned above. At the end of the task and once we approved their participation in our study, we flagged the participants by awarding an 'MTurk qualification' to ensure they are unable to take part in our follow-up studies.

4.2. *Results*

4.2.1. Selection of the system with the best performance

For the *reward-based* question, 12 of the 16 participants (75%) chose the system in the *animation* condition as the one with the best performance, 3 participants (19%) chose instead the system in the *no-animation* condition, while the remaining one indicated that both systems had the same performance level. Only 1 participant answered the *non-reward-based* question differently than the *reward-based* one, changing the answer from "animation" to "no-animation". These results are illustrated in Figure 5.

[7] `https://www.turkprime.com/Home/FrequentlyAskedQuestions`
[8] this information was displayed in the HITs' description

4.2.2. Reasons for choosing one system over the other – reward-based question

The responses to the question about why participants selected a particular HWR as the one the majority would choose to have the best performance were summarised through thematic analysis. Each response was associated with one or two of the following five themes: *number of errors, generic, type of error, animation,* and *others' opinion.* Figure 6 illustrates the frequencies of these themes. The categories *number of errors, generic,* and *type of errors* were the same as Study 1. The category *animation* was used when comments were related to the animation, e.g.: "Actually seeing the words transcribed probably leaves a good impression." Finally, the response of one participant that selected the HWR of the *animation* condition based on what other participants would select ("I think the second works better because the workers are more prepared at that point.") was categorised as *others' opinion.*

4.2.3. Reasons for choosing one system over the other – non-reward-based question

Thematic analysis was also applied to the answers related to the non-reward based question. The same themes as the previous question emerged, their frequencies are reported in Figure 7. For the majority of participants, 13 out of 16, the answers were the same (in terms of themes) as for the reward-based question. Only 3 participants answered this question differently than the previous one, and of these 3 only 1 changed their selection. In particular, the participant who selected a different system commented that they believed that other participants will choose the system in the *animation* condition because of the animation itself ("Actually seeing the words transcribed probably leaves a good impression."). Regarding the other 2 participants who provided different reasons without changing their selection, in one case the answer went from "type of errors and number of errors" to just "number of errors", on the other case, it went from "number of errors" to "animation" ("I liked that the second computer program highlighted the text in red as it was transcribing it").

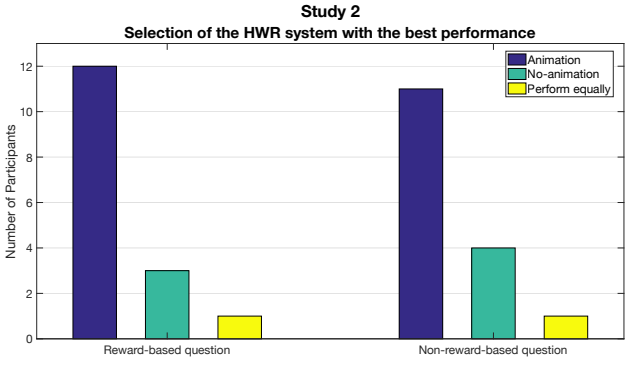

Figure 5. Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 2. Number of participants on the y-axis.

4.2.4. Performance ratings

Furthermore, a Wilcoxon Signed-rank Test revealed that the performance evaluation was higher in the *animation* condition ($Mdn = 5$) than in the *no-animation* condition ($Mdn = 4$), ($Z = 2.45, p < .05, r = 0.43$). Figure 8 shows participants evaluation of the performance of the systems.

4.3. *Discussion*

The results of this study clearly showed that the positive effect of animation cues persist even in a less controlled environment. The majority of participants reported that the system which integrated the animation performed better than the other system,

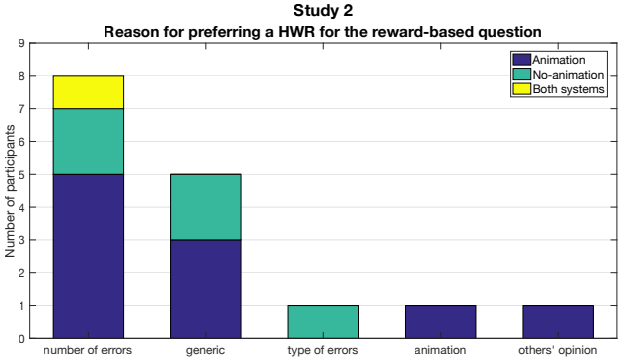

Figure 6. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

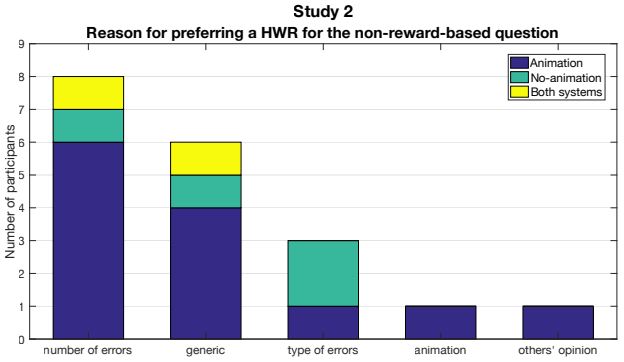

Figure 7. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the non-reward-based question. Number of participants on the y-axis.

for both the reward-based and non-reward-based questions. The statistically significant differences in the Likert scales results, as well as the qualitative data further confirm this finding. Moreover, similar to those from Study 1, the level of detail of the responses we collected clearly shows that the participants were committed to the task, giving credibility to the data. For example, participants referred not only to the number of errors that they found in the transcribed text but also to the type of errors (e.g., "I think the first program had more problems distinguishing the 'a' and 'o'."). Only one participant answered differently between the reward-based and the non-reward-based questions. This finding suggests that answers were not solely motivated by the financial incentives.

Only two participants justified their selection in terms of others' opinions or impressions, and by explicitly referring to the animation. This finding can be interpreted as confirming that the effect of motion cues is mostly unconscious. It should also be noted that these two participants did not change their answer between the reward-based and non-reward-based questions. This appears counterintuitive, and it can perhaps be explained as these two participants not paying much attention to the non-reward-based question. It should be noted that these references to others' opinions and to the animation emerged in Study 2, but not in Study 1. Such difference can be explained by the less controlled nature of Study 2.

The alignment of the results from Studies 1 and 2 also indicates that to further investigate this phenomenon, follow-up studies can be conducted on the MTurk platform, with considerable practical advantages. Having observed the effect of animation cues in a less controlled crowdsourcing environment, we turn to investigating whether this effect is specific to the handwriting recognition system we used so far, or whether the results can be generalized to a different type of smart system.

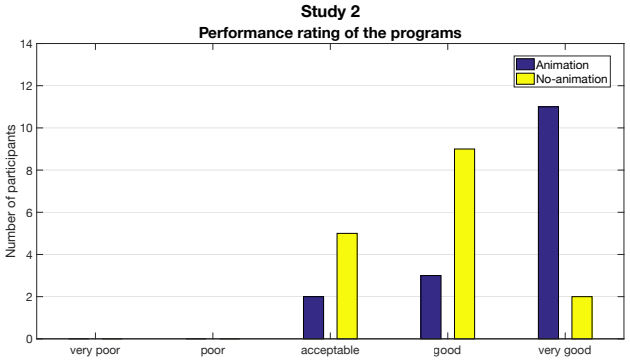

Figure 8. Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

5. Study 3

Studies 1 and 2 tested the effect of animation cues using one particular system, a HWR system. Handwriting recognition is in its very nature a visual task, making us wonder whether this factor alone may explain our results. So we designed a new study to assess whether the same effect would occur with a different type of system, and one involving processing that is not visual in nature. We selected a part-of-speech (POS) tagging system, a system that analyses natural language sentences and tags each word according to its syntactic function, such as article, adjective, adverb, conjunctions, noun, preposition, pronoun, and verb. POS tagging algorithms are readily available through open source libraries[9] and their application has been suggested for different types of interfaces and visualizations (Chuang, Manning, & Heer (2012); Yatani, Novati, Trusty, & Truong (2011)). We decided to continue to use text as the type of data handled by the smart system, for continuity with the previous studies and hence facilitate comparison of the results.

5.1. *Method*

5.1.1. Study Design

The study design was almost identical to Study 2: fully counterbalanced, within-participants, where participants were asked to evaluate and compare the performance of two systems, each corresponding to an experimental condition: *animation* and *no-animation*. There were only two differences. First, the two conditions were applied to a POS tagging system on a piece of text in English, rather than an HWR one (in Filipino), as illustrated in Figure 9. Specifically, in this study the animation shows the tags of the two last sentences appearing a few milliseconds one after the other next to each of the corresponding words[10]. Second, interpreting the results of the POS tagging system requires familiarity with POS tagging as a grammatical exercise. Because not everyone might be familiar with this, we included a *validation task*: participants had to tag a given sentence (in English) with the POS corresponding to each word. Only those who completed this validation task with less than 3 mistakes (out of 8 words) were allowed to proceed to the main task. Such validation task was not necessary for the HWR system, because anyone could complete that one by visual inspection.

The reward structure was also identical to Study 2, with a fixed amount being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority. To account for the slightly longer duration of the study (1 extra minute, due to the validation task) the reward for Study 3 was $1.33.

5.1.2. Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 2. The sample size was again 16, reported age ranged from 21 to 54 ($M = 26, SD = 10.02$), 5 males (31%) and 11 females (69%). 12 of them were United States nationals, 1 South Korean, 1 Canadian, 1 Belgian and 1 Bangladeshi national. The education levels of the participants ranged from primary school level to doctoral degree level or equivalent. Overall, 9 participants had a university degree, 4 completed secondary school, and 3 completed primary school.

[9] `http://www.nltk.org/`.
[10] `https://vimeo.com/210299892`

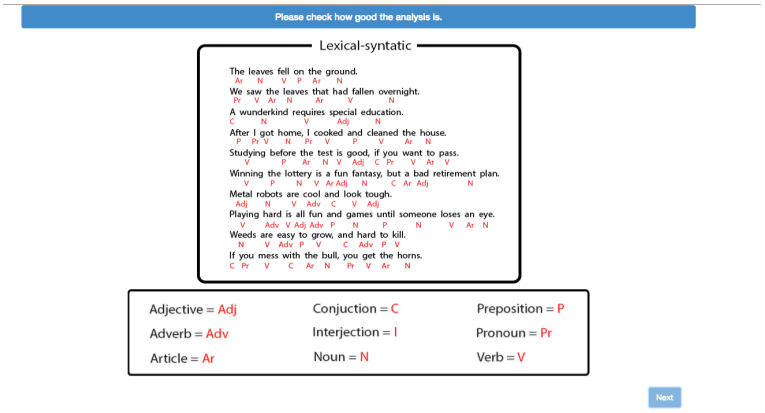

Figure 9. User Interface of the Part-of-Speech tagging system implemented in the Study 3

5.1.3. Equipment

The Web application used for Studies 1 and 2 was modified to include the validation task described above, and to show the POS tagging system in place of the HWR one. As in Study 2, the Web application was deployed to a publicly accessible Web server, to allow MTurk' workers to access it from their personal computers.

5.1.4. Procedure

This study followed the same procedure as Study 2, with the exception of the additional validation task outlined above. The validation task was displayed after the initial questionnaire about demographic information and before the main task. In the validation task participants were shown a POS-tagged sentence as an example. Then they were asked to tag one sentence. As mentioned above, if participants made less than 3 mistakes in this exercise they proceeded to the comparison of the two experimental systems, as in Study 2.

5.2. *Results*

5.2.1. Selection of system with the best performance

For the *reward-based* question, 11 of the 16 participants (69%) selected the system in the *animation* condition as the one with the best performance, 2 participants (12%) selected instead the system in the *no-animation* condition, while the remaining 3 (19%) expressed that both systems had the same performance. Moreover, one participant changed her choice for the *non-reward-based* question from "both systems" to "animation". The results are illustrated in Figure 10.

5.2.2. Reasons for choosing one system over the other – reward-based question

The responses to the question about why participants selected a particular POS as the one the majority will choose as the best performing system were summarised through thematic analysis. Each response was associated with one or two of the following three themes: *number of errors, generic* and *type of errors*. Figure 11 illustrates the frequencies of these themes. The themes were similar to those emerged from previous studies, with the exception that "type of errors" referred to specific POS tagging errors ("It seemed to have less mistakes. For example, the first one called "that" an article").

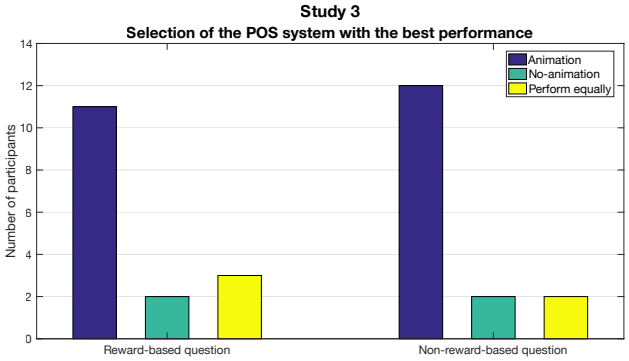

Figure 10. Selection of the participants for preferring a POS for the reward-based and non-reward-based questions in Study 3. Number of participants on the y-axis.

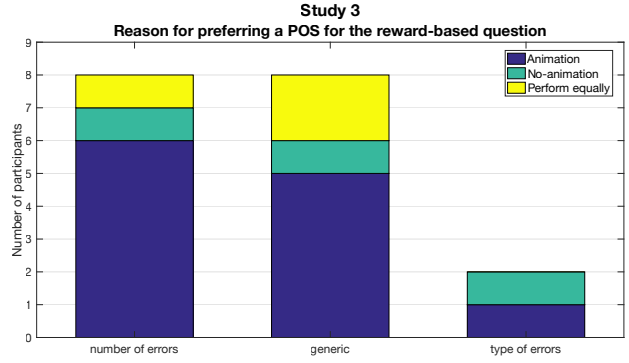

Figure 11. Reasons expressed by participants for preferring a POS in the *animation* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

5.2.3. Reasons for choosing one system over the other – non-reward-based question

Participants' responses about why they personally considered a particular POS system as the best performing one or the systems had the same performance were summarised through thematic analysis. Each response was associated with one or more of the following three themes: *generic, number of errors,* and *random*. Figure 12 illustrates the frequencies of these themes. The themes were similar to those which emerged from previous studies.

For the majority of participants, 12 out of 16, the answers were the same (in terms of themes) as for the reward-based question. Only 4 participants answered this question differently than the previous one, and of these 4 only 1 changed their selection. In particular, the participant who selected a different system commented that she considered that both had the same performance, but she left a comment that her selection was random (e.g. "It's really a toss up. I saw the same potential errors on the same word in both programs, so I'm just picking one."). Regarding the other 3 participants who provided different reasons without changing their selection, in one case the answer went from "type of error" to just "generic", on the second case went from "number of errors" to "number of errors and generic", in the last case, it went from "number of errors and type of errors" to just "number for errors".

5.2.4. Performance ratings

A Wilcoxon Signed-rank Test revealed that the performance evaluation was higher in the *animation* condition ($Mdn = 4$) than in the *no-animation* condition ($Mdn = 3$), ($Z = 2.55, p < 0.05, r = 0.45$). Figure 13 shows participants evaluation of the performance of the systems.

5.3. *Discussion*

The results of Study 3 extend those of Study 2, they demonstrate that the effect of animation cues on participants' perception of system performance applies also to the

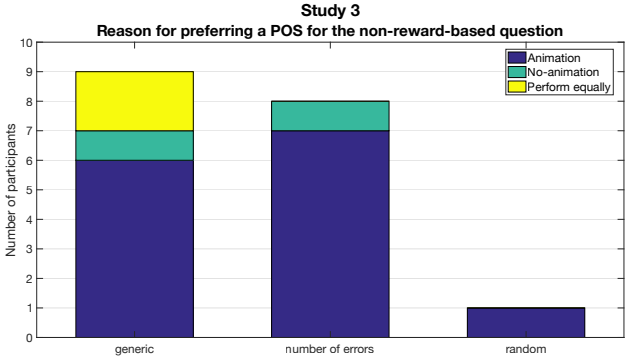

Figure 12. Reasons expressed by participants for preferring a POS in the *animation* or *no-animation* condition for the non-reward-based question. Number of participants on the y-axis.

POS tagging system we tested. The majority of participants selected the system in the *animation* condition as the one with the best performance, and the Likert-scale ratings for this system were higher, in aggregate, than those for the *no-animation* condition, with statistical significance. Similar to Studies 1 and 2, the qualitative data collected in Study 3 indicates that participants offered a variety of reasons to justify their selections, and only 1 participant provided different answers based on the financial incentives, suggesting that most answers were not based solely on the financial incentives. Moreover, the themes emerged from the qualitative data are the same as Studies 1 and 2, further confirming the similarity of the effect on POS and on HWR systems. Such effect, then, appears to apply even if the task performed by the system is not inherently visual, and hence if the animation does not directly mimic the task performed by the smart system.

Having observed this effect both in the lab and on MTurk, and on two different systems, we turn to the question of *why* such effect occurs. Given that both animations highlight one word at a time, in reading order (from left to right), one option could be that the animations give users the impression that the systems process text in the same way a person would process it. Is it possible that the similarity to humans may positively influence users' attitude towards the system? This, in turn, may lead them to evaluate its performance more favourably, perhaps somehow suggesting to them that the system is "as smart as a person". An alternative explanation might involve more generally the relationship between the animations in our studies and users' mental model of how the system works (Balijepally, Nerur, & Mahapatra (2015)). In more detail, Balijepally et al. found that the quality of a developer's mental model positively impacted the users' performance as measured in terms of software quality. Moreover, the accuracy of a person's mental model of a system is based on the person's prediction of a system's future behaviour and therefore influences how they interact with it (Norman (2013)). The animations might induce a mental model that leads them to rate the system performance more positively. To assess the validity of these possible explanations, we designed and carried out three follow-up studies that we report in the following.

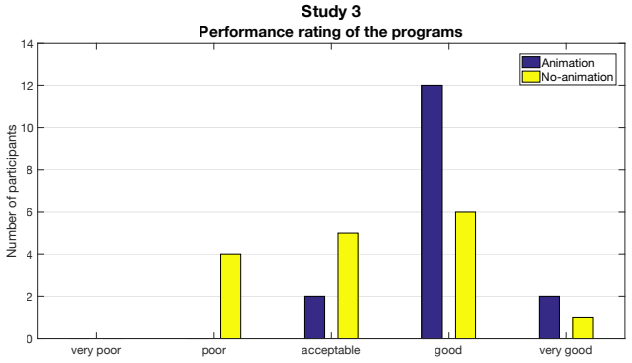

Figure 13. Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

6. Study 4

Through Studies 1, 2 and 3 we found that animation cues can influence how people evaluate the performance of smart systems. Why do these animations cues have an effect on participants' perception of the smart system's performance? Could the effect be due to the animations making the systems appear to process information like a human? If this is the case, then showing an animation where the order in which the words are processed is decisively not human-like should have no effect on participants' perception of the system performance. So we designed a fourth study to test whether an animation that is decisively not human-like would still cause the same effect as the animation used in the previous studies.

Study 3 revealed that the animation effect applies in a similar way to a POS tagging system as it does to a HWR system. For simplicity, then, we decided to conduct further experiments on the HWR system, as it does not require the additional training and validation task described above.

6.1. *Method*

6.1.1. Study Design

The study design was identical to Study 2: fully counterbalanced, within-participants, where participants were asked to evaluate and compare the performance of two systems, each corresponding to an experimental condition: *non-human-like animation* (*NHL-animation*, for short) and *no-animation*. The only difference was the animation: instead of highlighting words in left-to-right order (as in Studies 1, 2 and 3), the order was random[11].

The reward structure and amounts were also identical to Study 2, with a fixed amount of $ 1.17 being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority.

6.1.2. Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 2. The sample size was again 16, age ranged from 22 to 44 ($M = 31, SD = 7.16$), 13 males (81%) and 3 females (19%). All except for one of the participants reported to be United States nationals, and the remaining one German. The education levels of the participants ranged from primary school level to university degree level or equivalent. Overall, 7 participants had a university degree, 7 completed secondary school, and 2 completed primary school. None of the participants reported knowing Filipino.

6.1.3. Equipment

The same Web application used for tudies 1 and 2 was used for Study 4, with the only difference of the animation, as described above.

6.1.4. Procedure

The procedure was the same as Study 2.

[11] `https://vimeo.com/183550733`

6.2. *Results*

6.2.1. Selection of the system with the best performance

For the *reward-based* question, 10 of the 16 participants (62%) selected the system in the *NHL-animation* condition as the one with the best performance. The other 3 participants (19%) indicated that the system in the *no-animation* condition performed the best, while the remaining 3 (19%) suggested that both systems had the same performance. None of the participants changed their choice for the *non-reward-based* question. The results are illustrated in Figure 14.

6.2.2. Reasons for choosing one system over the other – reward-based question

The responses to the question about why participants selected a particular HWR as the one the majority will choose as the best performing one were summarised through thematic analysis. Each response was associated with one or two of the following five themes: *number of errors, generic, type of errors, speed,* and *animation*. Figure 15 illustrates the frequencies of these themes. All themes except for *speed* were the same as in previous studies. Responses categorised as *speed* are related to comments when participants mentioned that the speed of the system was a reason for their choice. One example of these responses is "Seems that the first program was faster and presented a complete page at once."

6.2.3. Reasons for choosing one system over the other – non-reward-based question.

Thematic analysis was also applied to the answers related to the non-reward based question. Each response was associated with one or two of the following six themes: *generic, number of errors, speed, type of errors, animation,* and *others' opinion*. Figure 16 illustrates the frequencies of these themes. The themes were similar to those which emerged from previous studies. For the majority of participants, 10 out of 16, the answers were the same (in terms of themes) as for the reward-based question. Only

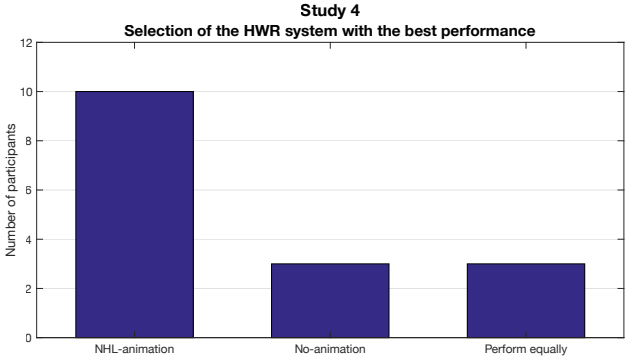

Figure 14. Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 4. Number of participants on the y-axis.

6 participants answered this question differently than the previous one. In particular, the participants who provided different reasons without changing their selection, in one case the answer went from "type of errors and number of errors" to just "generic", on the second case went from "number of errors" to "type of errors". The third case change from "animation" to "generic", in the next case went from "number of errors" to "generic". The fifth case his answer was categorised first as "speed" and changed to "other's opinion". Finally, in the last case went from "type of errors" to just "speed".

6.2.4. Performance ratings

A Wilcoxon Signed-rank Test revealed that the performance evaluation was higher in the *NHL-animation* condition ($Mdn = 5$) than in the *no-animation* condition

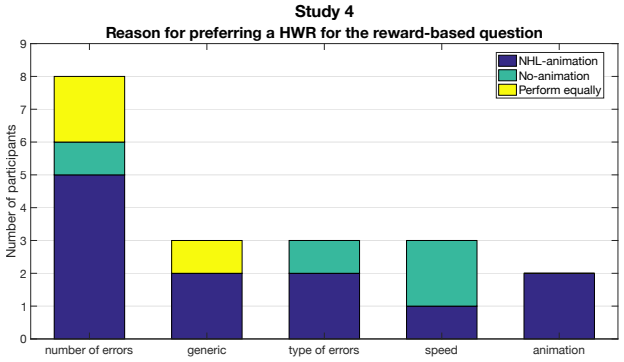

Figure 15. Reasons expressed by participants for preferring a HWR in the *NHL-animation* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

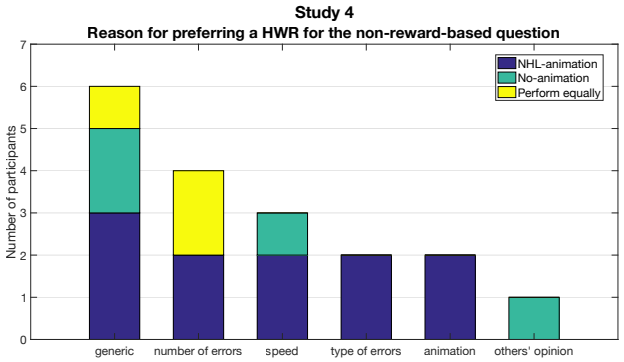

Figure 16. Reasons expressed by participants for preferring a HWR in the *NHL-animation* or no-anima*no-animation*tion condition for the non-reward-based question. Number of participants on the y-axis.

($Mdn = 4$), ($Z = 2.07, p < 0.05, r = 0.37$). Figure 17 shows participants evaluation of the performance of the systems.

6.3. *Discussion*

The majority of participants in Study 4 selected the system in the *NHL-animation* condition as the one with the best performance, and the Likert-scale ratings for this system were higher, in aggregate, than those for the *no-animation* condition, with statistical significance. The analysis of qualitative data is also very much in line with that of previous studies. These results indicate that the effect we observed in previous studies can be observed also for animations that can be interpreted as non-human-like. Therefore the tentative explanation suggested above, that the effect of animations in Studies 1 to 3 may be related to making the system appear more human-like can be rejected. Having rejected human-like explanation, in what follows we turn to the option that the effect of animations may be due to the more general relationship between the animation and a user's mental model of the smart system.

7. Study 5

A new study was designed to investigate the relationship between users' mental models of HWR systems and the animations we displayed in earlier studies. In particular, in this study participants were asked one open question about their idea of how an HWR system works, to check whether these explanations are compatible with the animations used in our prior studies. The study then followed the same structure as Study 2, but at the end we also asked participants whether their experience of using the system matched their initial idea of how it works.

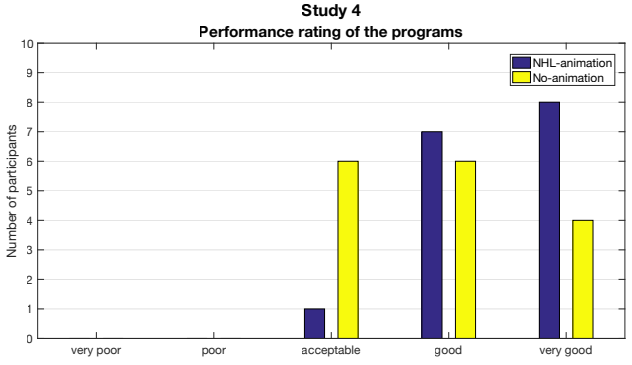

Figure 17. Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

7.1. *Method*

7.1.1. Study Design

The study design was almost identical to Study 2: fully counterbalanced, within-participants, where participants were asked to evaluate and compare the performance of two systems, each corresponding to an experimental condition: *animation* and *no-animation*. The only difference was that we included two additional questions mentioned above.

The reward structure was also identical to Study 2, with a fixed amount being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority. To account for the slightly longer duration of the study (1 extra minute, due to the additional question) the reward for Study 5 was $1.33.

7.1.2. Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 2. The sample size was again 16, reported age ranged from 22 to 37 ($M = 29.5, SD = 4.76$), 11 males (69%) and 5 females (31%). All except one were United States nationals, with the remaining 1 South Korean. The education levels of the participants ranged from secondary school level to masters degree level or equivalent. Overall, 2 participants had a master's degree, 9 a university degree, and 5 completed secondary school. None of the participants reported knowing Filipino.

7.1.3. Equipment

The same Web application used for Study 2 was used for Study 4, with the only addition of the extra questions, as described above.

7.1.4. Procedure

This study followed the same procedure as Study 2, with the exception of the additional open question about how the system works, as described above. The additional question was asked after the initial questionnaire about demographic information and before the main task. In an attempt to prevent random answers, participants were required to submit answers containing at least 20 words. After answering this question participants proceeded to comparing the two conditions as in Study 2.

7.2. *Results*

7.2.1. How people think a HWR works

The responses to the question regarding how participants think that the HWRs work were analysed through thematic analysis. Two themes emerged in our analysis: *match with database* and *image recognition*. The theme *match with database* included responses that mention using a database to compare the words or characters identified in the handwritten text, such as "The program analyses the written text. It then compares each character to a database loaded into it [...]". The theme *image recognition* was associated to responses that mention how the program processes images to extract characters and words, such as: "It scans the handwriting into an image and then the program look[s] at the image pixel by pixel to match each individual letter [...]'.

Participants' answers suggest that the majority of them seem to have a shared mental model of how the system works. In general, participants agree that somehow the system has to detect the words or letters to digitise them. Of the 16 participants, 11 stated that the system matches the words and letters using some form of optical recognition. Furthermore, 11 participants mentioned that this then needs to be matched with a 'collection' of some kind, containing labelled examples of handwritten text, to find the corresponding letter or word.

7.2.2. Selection of the system with the best performance

For the *reward-based* question, 5 of the 16 participants (31%) chose the system in the *animation* condition as the one with the best performance. Additionally, 7 participants (44%) selected the system in the *no-animation* condition as the one with the best performance, while the remaining four (25%) indicated that both systems had the same performance level. Only 1 participant answered the *non-reward-based* question differently than the *reward-based* one, changing the answer from "both systems" to "no-animation". These results are illustrated in Figure 18.

7.2.3. Reasons for choosing one system over the other – reward-based question

The responses to the question about why participants selected a particular HWR as the one the majority would choose to have the best performance were summarised through thematic analysis. Each response was associated with one or two of the following five themes: *generic, number of errors, type of errors, animation, speed* and *others' opinion*. Figure 19 illustrates the frequencies of these themes. The themes definitions were the same as in prior studies.

7.2.4. Reasons for choosing one system over the other – non-reward-based question

We also analysed why participants considered a particular HWR as the one with the best performance for themselves and concluded with five themes in total: *number of*

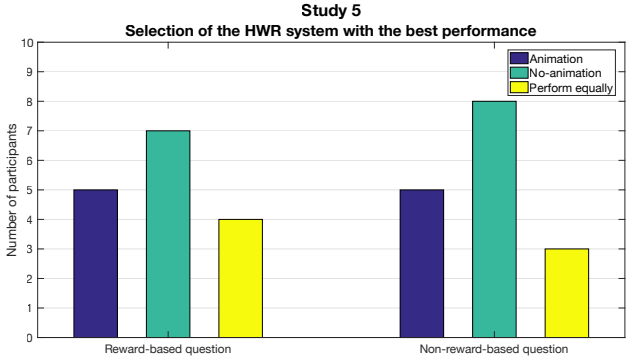

Figure 18. Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 5. Number of participants on the y-axis.

errors, generic, type of errors, animation, and *speed*. The five themes are the same as above. Figure 20 illustrates the frequencies of these themes.

For the majority of participants, 9 out of 16, the answers were the same (in terms of themes) as for the reward-based question. Only 7 participants answered this question differently than the previous one, and of these 7 only 3 changed their selection. In particular, the first participant who selected a different system commented that she believed that other participants would choose the system in the *no-animation* condition because it had the best performance ("I thought the second showed actual better performance"). The second participant considered the majority would select the *animation* condition because this person felt that the other participants would like to see how the system is working. Finally, the last participant mentioned that others would not see a difference between both systems, as such, he selected *both systems* performed equally for the reward-based question and changed to *no-animation* condition for the non-reward-based question. Regarding the other 2 participants who provided different reasons without changing their selection, in one case the answer went from "type of errors and number of errors" to just "number of errors", on the other case, it went from "number of errors" to "animation" ("I liked that the second computer program highlighted the text in red as it was transcribing it").

7.2.5. Performance ratings

A Wilcoxon Signed-rank Test revealed no significant difference in the evaluation of the performance between the *animation* condition and *no-animation* condition ($Z = 1.03, p = 0.31, r = 0.18$). Figure 21 shows participants evaluation of the performance of the systems.

7.2.6. The system worked as participants expected.

All participants reported that both systems successfully transcribed the handwritten text to typed text, and so they considered that the systems worked as they expected. Additionally, only three participants mentioned in their comments the animation (e.g.

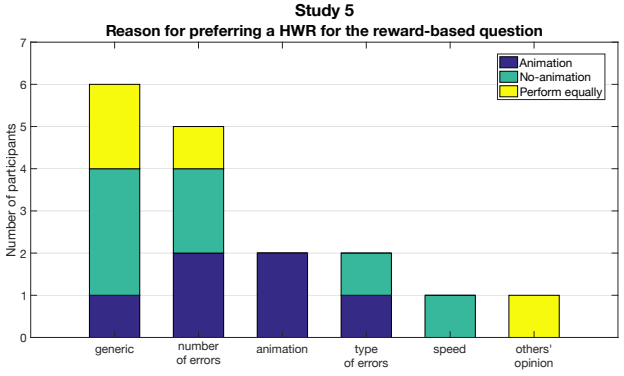

Figure 19. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

"For the second program, it showed how the program scanned each word in red. It was computing for the e-text").

7.3. *Discussion*

The explanations that participants provided about how an HWR system works matched quite closely how this kind of systems are actually implemented: they perform some form of image recognition on characters and words. All participants seem to have an accurate mental model regardless of whether they have formal technical background (indeed only 4 of them did). Moreover, the explanations participants provided seem to be quite in line with the animation that we implemented for Studies 1,

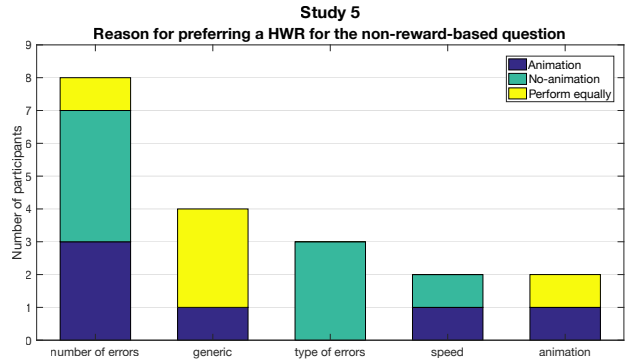

Figure 20. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the non-reward-based question. Number of participants on the y-axis.

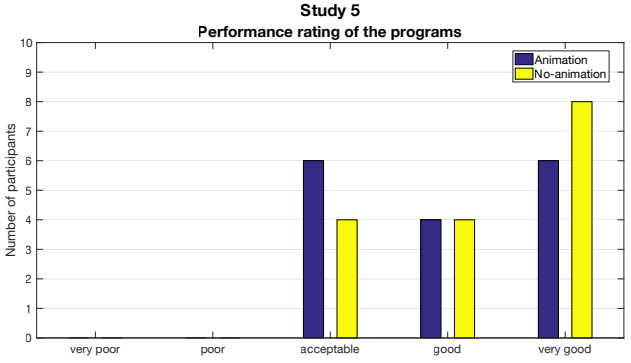

Figure 21. Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

2 and (to an extent) 4. The match, however, is not always an exact one: 14 out of the 16 participants explained that the recognition would happen letter by letter, so in the same way a human would actually type handwritten text into a computer. In contrast, the animation implemented in the previous studies can be interpreted as processing the text word by word rather than character by character.

The results from Study 5 seem to be in stark contrast to those from Studies 1 to 4. Only 5 participants selected the system in the *animation* condition as the one with the best performance level, and the analysis of the Likert-scale ratings did not reveal statistically significant differences between the conditions, despite the sample size being the same as in the earlier studies. The different results can be attributed to the additional question about participants' mental model of HWR systems asked at the beginning of the study.

Arguably, asking participants how they think an HWR system works makes their mental model for this kind of system salient to them. This salience seems to contrast the effect of the animation that we observed in earlier studies. Perhaps, then, making participants aware of how the system works has an effect similar to that of the animation in our earlier studies. In other words, these results suggest that in our earlier studies the animations reminded participants of how the smart system works, instead in Study 5 the preliminary question had the same effect, so it seems to have replaced the effect of the animation (for both conditions). This finding resonates with studies in Psychology which demonstrated that making a bias salient to participants may remove the effect of the bias. In particular Schwarz & Clore (1983) through a well known study about the effect of weather on mood demonstrated that asking participants about the weather (and hence making the weather salient to them) removes the effect that weather has on mood (at least in the case of bad weather). Similarly, in our study asking participants about how the system works seems to remove the effect of the animation. We further explore the relationship between mental models and the effect of animations on perceptions of performance in the following study.

8. Study 6

The results of Study 5 suggest that the animation used in Studies 1 and 2, and to some extent in Study 4, matched participants' mental models of HWR systems. Based on such finding, a possible explanation of the results from Studies 1, 2 and 4 is that the animations we displayed "reassured" participants that they system work as they expected. As such, the animation raised their confidence in the system and enhanced their perception of its performance.

Based on this, we formulated an explanation of how a HWR system works which matches the animation that we used in Studies 1 and 2. We refer to this one as the *original* animation. We also designed a new animation, which we refer to as the *alternative* animation, and we formulated a corresponding explanation. The *alternative* animation consisted of enclosing each word with a rectangle and inverting its colour, before displaying the corresponding word on the right hand side of the screen[12]. This alternative animation was designed to be at odds with the explanations collected from participants in Study 5 about how a HWR works. For consistency, both animations, original and alternative, included the same transcription errors. Study 6 was designed to compare and contrast the effect of *matching* and *mismatching* animations and

[12] `https://vimeo.com/183480642`.

explanations.

8.1. *Method*

8.1.1. Study Design

The two animations and the two explanations described above define 4 conditions in a 2×2 fashion: (original animation, original explanation), (original animation, alternative explanation), (alternative animation, alternative explanation), and (alternative animation, original explanation). In the first and third condition, animation and explanation are *matching*, while in the second and fourth they are *mismatching*. These 4 conditions were applied through a between-participants study design: each participant was assigned to one of these 4 conditions. Similar to Study 2, each participant was asked to evaluate and compare the performance of two HWR systems: one involving an animation (*animation* condition) and one with no animation (*no-animation*). The *no-animation* condition, which was similar to the condition in Study 2, was the same for all participants. The *animation* condition would involve either the *original* or the *alternative* animation, depending on the assigned group of the participant. The *animation* and *no-animation* conditions were presented to participants in fully counterbalanced order.

The reward structure was identical to Study 2, with a fixed amount being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority. To account for the slightly longer duration of the study (compared to Study 2) the reward for Study 6 was $1.33.

8.1.2. Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 2. The sample size was 64, higher compared to earlier studies, to reflect the larger number of conditions and the between-participants design. Age ranged from 20 to 61 ($M = 32, SD = 11.22$), 34 males (53%) and 30 females (47%). All except for 2 of the participants reported to be United States nationals, the remaining two being Canadian and South Korean. The education levels of the participants ranged from secondary school level to doctoral degree: 1 participant had a doctoral degree, 5 had a masters' degree, 39 a university degree, and 19 completed secondary school. One of the participants reported knowing Filipino.

8.1.3. Equipment

The Web application used for Study 2 was modified to include the two different animations described above, and to include an initial explanation of how the system works, together with a *reinforcing question*, as detailed below. As in Study 2, the Web application was deployed to a publicly accessible Web server, to allow MTurk' workers to access it from their personal computers. Two additional questions were also included in the final questionnaire, to ask participants whether they considered that the systems worked according to the explanation they received at the beginning of the study and why.

8.1.4. Procedure

In addition to the procedure followed in Study 2, the explanation of how the system works (*original* or *alternative*, depending on the condition) was presented to the participants. The explanation was given to the participants after the questionnaire where we asked them about their personal information, with the aim of influencing their mental model of the system. However, participants were not explicitly told that the explanation would correspond to any animation cues. After the initial explanation, participants were asked to explain, in their own words, how they think the systems works – we refer to this as the *reinforcing question*. participants were told that their responses should have more than 20 words to be considered valid. Subsequently, the two HWR systems were presented to the participants. Similar to the previous studies, participants were then asked to evaluate the performance of the systems on a 5-point Likert scale, to select the system they considered to perform best (or that the systems had the same performance level) and to justify their selection. At the end of the study, after evaluating both systems, participants answered the two new questions that we designed.

8.2. *Results*

For the purpose of the analysis we group the results corresponding to the two conditions with *matching* explanation and animation –namely: (original animation, original explanation) and (alternative animation, alternative explanation)– and those corresponding to the two conditions with *mismatching* explanation and animation –namely: (original animation, alternative explanation) and (alternative animation, original explanation). This applies both to the quantitative and the qualitative analysis.

8.2.1. Selection of the system with the best performance

Matching conditions. For the *reward-based* question, 21 of the 32 participants (66%) selected the system in the *animation* condition as the one with the best performance. Additionally, 5 participants (15%) selected the system in the *no-animation* condition as the one with the best performance, while the remaining 6 (19%) indicated that both systems had the same performance level. None of the participants answered the *non-reward-based* question differently than the *reward-based* one. The participant who reported knowing Filipino was in the this condition, and she selected the system in the *animation* condition as the one with the best performance for both questions.

Mismatching conditions. For the *reward-based* question, 10 of the 32 participants (31%) selected the system in the *animation* condition as the one with the best performance. Additionally, 12 participants (38%) selected the system in the *no-animation* condition as the one with the best performance, while the remaining 10 (31%) indicated that both systems had the same performance level. 16 of the participants answered the *non-reward-based* question differently than the *reward-based* one.

8.2.2. Reasons for choosing one system over the other – reward-based question.

Participants' responses to the question about why they chose a particular HWR system as the one that the majority will choose to have the best performance were categorised through thematic analysis. Each response was associated with one or two themes, with five themes found in total: *number of errors, generic, animation, type of errors, speed* and *others' opinion*. The five themes are the same as in Study 2. Figure 22 illustrates

the frequencies of these themes, also classified on the reward-based question (*animation* or *no-animation*) for the **matching conditions**, while Figure 23 illustrates the frequencies for the **mismatching conditions**. The participant who reported knowing Filipino was in the matching condition, and her answers were associated with the theme *animation* for both questions.

8.2.3. Reasons for choosing one system over the other – non-reward-based question.

Participants' responses to the question why they selected a particular HWR as the one with the best performance were categorised through thematic analysis. Each response was associated with one or two themes, with five themes found in total: *number of errors, generic, type of errors, animation* and *speed*. The themes categorised participants'

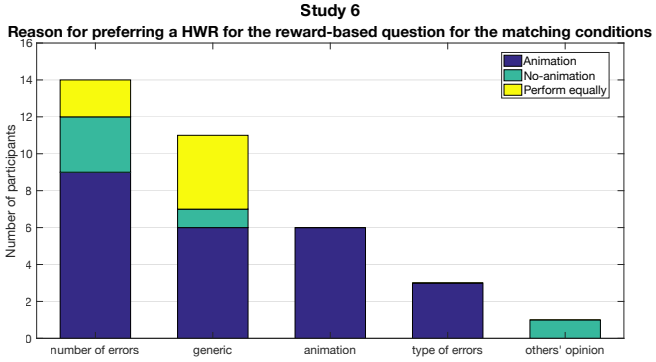

Figure 22. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question for the match conditions. Number of participants on the y-axis.

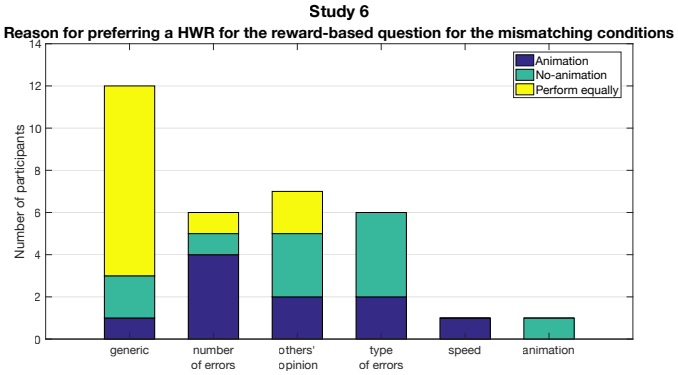

Figure 23. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question in the match conditions. Number of participants on the y-axis.

comments as we did in the previous studies. Figure 24 illustrates the frequencies of these themes, also classified on the individual preference (*animation* or *no-animation*) for the **matching conditions**, while Figure 25 illustrates the frequencies for the **mismatching conditions**.

8.2.4. System working according to expectations

Matching conditions. Overall, 27 of the 32 participants indicated that the systems worked as they expected from the explanation given at the beginning of the study, while the remaining 5 stated that it did not. In more detail, participants considered that the system compares each word with a database. As such, 3 participants believed the system would process the data word by word rather than character by character

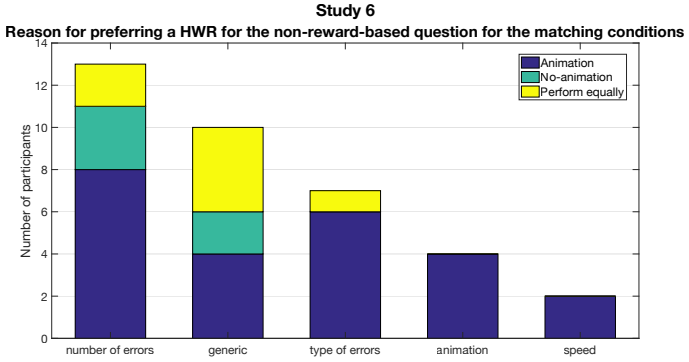

Figure 24. Reasons expressed by participants for preferring a HWR in the animation or *no-animation* condition for the non-reward-based question in the match conditions. Number of participants on the y-axis.

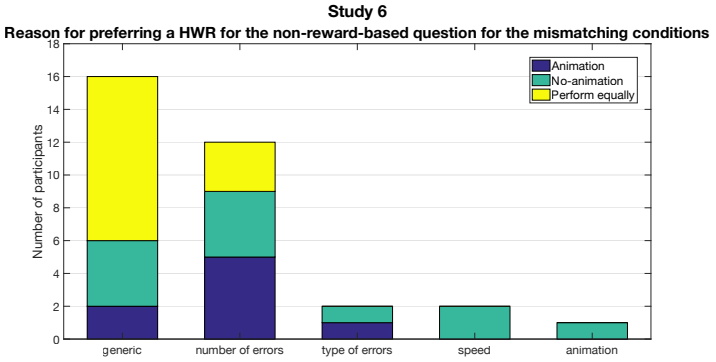

Figure 25. Reasons expressed by participants for preferring a HWR in the animation or *no-animation* condition for the non-reward-based question for the mismatch conditions. Number of participants on the y-axis.

as the errors they found (e.g. "It seemed the program did it letter by letter, not by the word as described. But then again I don't know the language, so changing one letter like the programs did may still have been recognizing a word."). Other 2 participants mentioned that they believe that the systems did not work because only one system show the animation (e.g. "The second computer program highlights the words as it transcribes them. The first didn't appear to do that."). In the free text comments, 10 participants mentioned the animation explicitly as a reason of why they considered the system worked as they expected (e.g. "I could see the text being highlighted and picked apart", and "Because you could see the process of transcription as it was happening.").

Mismatching conditions. Overall, 25 of the 32 participants indicated that the system worked as expected from the explanation given at the beginning of the study, while the remaining 7 stated it did not. In the free text comments, 10 participants mentioned the animation. In more detail, 3 participants mentioned that the animations mismatched the explanation. Moreover, the 3 participants who reported that the system did not work as they expected mentioned that they thought the system transcribed the handwritten text word by word rather than character by character because of errors they found in the typed text. In contrast, other participants felt that the system transcribed the handwritten text better than they expected. Because of this, they felt that the system worked correctly.

Participants' responses to why they considered the that the systems worked (or not) according to their expectations were categorised through thematic analysis. Each response was associated with one or two themes, with eight themes in total: *generic, animation, faith, disbelief, correctness, analysis, experience,* and *technology.* Figure 26 illustrates the frequencies of these themes for the **matching conditions**, while Figure 27 illustrates the frequencies for the **mismatching conditions**. The theme generic was associated to responses where participants did not provide full explanation or misunderstood the question, such as "I believe they translated the handwritten text into a digital computer font". The theme *animation* was used when participants talked about why the animation affected their consideration of whether the systems are working or not, such as: "Because you could see the process of transcription as it was happening". We grouped a response into the theme *faith* if it is related to the participants believing the explanation provided: "I had no reason to doubt the explanation, it seemed perfectly reasonable". Comments grouped into *disbelief* is the opposite, and instead it's related to situations where the participants do not believe in the explanation provided: "I don't see how changing the color of the handwritten text to match the color of the paper as a way to convert the text to etext [...]". Responses related to the accuracy of the output, such as "Yes, it translated the characters of the handwritten text correctly", were categorised as *correctness.* The theme *analysis* was used to categorise comments that talk about the actual transcription process, such as "It appears that the program goes through each letter and tries to identify which letter it is". The theme *experience* was used for any comments in which the participant talk about his/her own personal experience with HWR systems: "I've used OCR [Optical Character Recognition] programs before and they were never as accurate as this one was, so I don't believe it actually exists". Finally comments such as "Technology and artificial intelligence is growing at an exponential rate" were categorised as *technology.*

8.2.5. Performance ratings

Matching conditions. A Wilcoxon Signed-ranks Test revealed that the performance evaluation was higher for *animation* condition ($Mdn = 5$) than for *no-animation*

condition ($Mdn = 4$), ($Z = 2.94, p < 0.01, r = 0.37$). Figure 28 shows participants evaluation of the performance of the systems.

Mismatching conditions. A Wilcoxon Signed-ranks Test did not reveal statistically significant differences. Figure 29 shows participants evaluation of the performance of the systems.

8.3. *Discussion*

The results of this study confirms what was suggested by the findings of Study 5: animation cues influence the perception of the system performance only if they are compatible with the participants mental model. More in general, taken together with

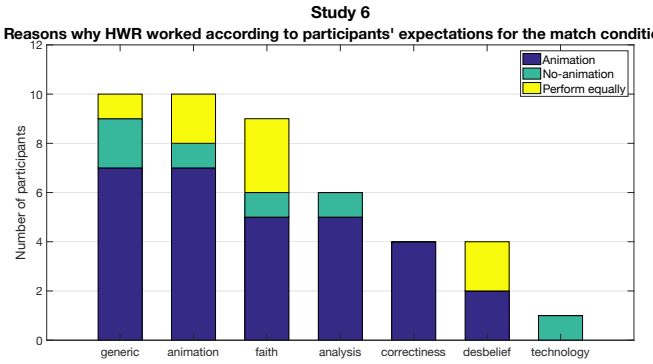

Figure 26. Reasons expressed by participants for why they considered that the systems worked according to their expectations for the match conditions. Number of participants on the y-axis.

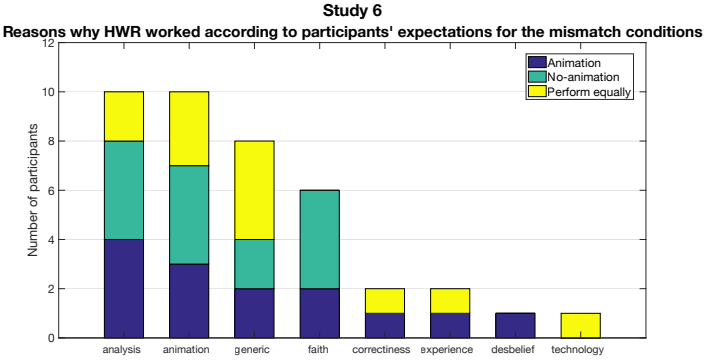

Figure 27. Reasons expressed by participants for why they considered that the systems worked according to their expectations for the mismatch conditions. Number of participants on the y-axis.

the results of Study 5, these results allow us to propose the following explanation for the effect:

> Animation cues suggest the user an explanation of how the smart system works. If this explanation is largely compatible with the user own mental model of the system, i.e. if this explanation appears plausible to the user, then they feel reassured about how the system functions, and therefore are inclined to have higher confidence in the system output, compared to an alternative system for which they have no cues about how it may work.

We used the word 'largely' because the results of Study 5, indicate that the explanations for how the system works provided by participants do not match *exactly* the animation. However, if the explanation is radically different, as in the *mismatching*

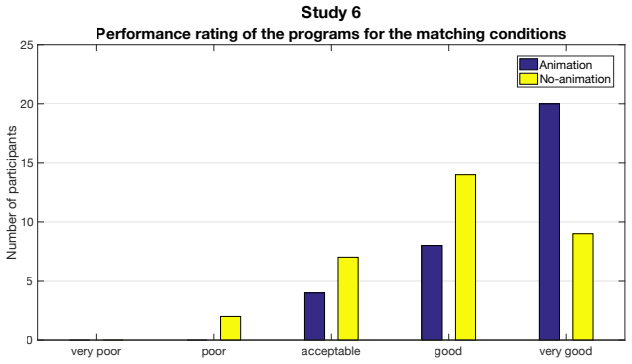

Figure 28. Likert-scale of participants' evaluation of the performance of the systems in the *animation* and *no-animation* conditions in the matching conditions. Number of participants on the y-axis.

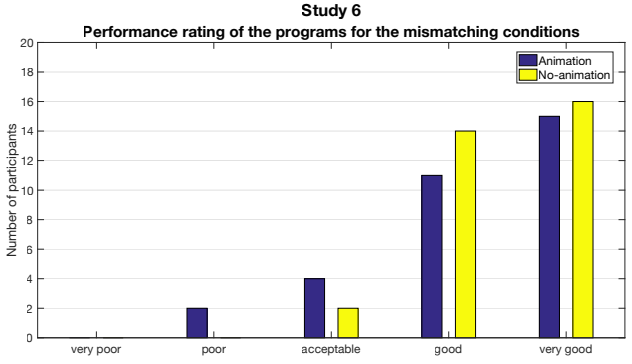

Figure 29. Likert-scale of participants' evaluation of the performance of the systems in the *animation* and *no-animation* conditions in the mismatching conditions. Number of participants on the y-axis.

conditions of Study 6, the effect disappears.

Moreover, as mentioned above, the results of Studies 1 to 4 suggest that this effect takes place largely unconsciously – most participants did not mention the animation in their justification for the selection of the system with the best performance. Similarly, in Study 5 we observed that if the explanation of how the system works is made salient to participants, in that case through an initial question, the effect of the animation disappears. Arguably, the explanations that participants provided in Study 5 could apply to *both* the systems they evaluated, so that process reminded them, or made them aware, of how *both* systems work. Hence, their judgement was not biased towards either of them. It should be noted that in Study 6 participants were also exposed to an explanation of how an HWR system works at the beginning of the study. However, in that case the explanation was provided to them, and it matches very closely the animation show – these differences may explain why the effect of animation was still observed in this study compared to Study 5.

In fact, 13 participants (out of 64) mentioned that the reason they think that the system works (or not) is largely based on the explanations they received.

Even though the importance of mental models in HCI has been discussed for at least three decades (Kieras & Bovair (1984); Norman (2013)), most prior work focussed on the effect of mental models on *users' performance* when using an interactive system. In contrast, our results suggest a relationship between mental models and users' perception of the *system's performance*.

In the qualitative data, we did not find comments of people explaining that they perceived a match or mismatch between the explanation that elicits a mental model and the animation they received. This behaviour suggests that the effect of animation cues happens unconsciously. Thus, we can argue that the participants' comments and evaluation make visible that indeed the fact that the animation matches participants' mental model affects their perception. The participant who reported knowing Filipino selected the same answers as the majority of other participants (the system in the *animation* condition as the one working better), so her knowledge of the language does not seem to play a visible role here. She justified her selection referring to the animation – perhaps her language knowledge allowed her to give more attention to this feature of the UI, compared to other participants who might have busy comparing the input and output portions of the interface. However, further work would be required to evaluate the effect of language familiarity on the effect of animation. Now that we found the reason behind why the *animation* cues influence people's perception on how they perceive smart systems' performance, we move to further characterize this effect with the following studies.

9. Study 7

Through Study 1 and 2, we found that animation cues can influence how people evaluate the performance of screen-based systems. As a subsequent step, we evaluate whether the amount of detail of a displayed motion, so how much animation needs to be shown in all the elements related to the system's task (e.g. handwritten text and e-text), can have an impact on participants' perception. We expect to find a relationship between the amount of motion displayed and the perceived performance.

To explore this, we designed a new animation, which involves less motion than the animations used in previous studies.

9.1. *Method*

9.1.1. Study Design

The study design was almost identical to Study 2: fully counterbalanced, within-participants. However, it involved 3 conditions: *animation*, *partial-animation*, and *no-animation*. Similar to prior studies, each condition corresponded to a system that participants were asked to evaluate and compare in terms of performance. The new *partial-animation* condition is similar to the *animation* condition, except that instead of involving the animation on both the input and output parts of the UI (i.e. on both the handwritten and typed text), it only applies to the output part of the UI (approximately the right half of the screen), while the input part of the UI remains static.

The reward structure was also identical to Study 2, with a fixed amount being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority. To account for the slightly longer duration of the study (compared to Study 2) the reward for Study 7 was $2.

9.1.2. Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 2. The sample size was 48, to account for the increased number of conditions, reported age ranged from 22 to 44 ($M = 34, SD = 9.53$), 31 males (65%) and 15 females (31%). Of those participants, 46 were United States nationals, 1 was Polish and the other was British. The education levels of the participants ranged from primary school level to masters' degree level or equivalent. Overall, 2 participants had a master's degree, 30 a university degree, 15 completed secondary school, and 1 completed primary school. Two of the participants reported knowing Filipino.

9.1.3. Equipment

The Web application used for Study 2 was modified to include the additional condition described above.

9.1.4. Procedure

This study followed the same procedure as Study 2, with the exception of the additional condition outlined above.

9.2. *Results*

9.2.1. Selection of the system with the best performance

For the *reward-based* question, 26 of the 48 participants (54%) selected the system in the *animation* condition as the one with the best performance, 9 participants (19%) selected instead the system in the *partial-animation* condition, 7 participants (15%) the system in the *no-motion* condition, while the remaining 6 participants (12%) suggested that the three systems had the same performance. Moreover, 5 participants changed their choices for the *non-reward-based* question as follows: from 'Partial-animation' to 'Animation', from 'All performed equally' to 'Animation', from 'Animation' to 'No-animation', from 'All performed equally' to 'No-animation', and from 'Animation' to

'All performed equally'. These results are illustrated in Figure 30.

9.2.2. Reasons for choosing one system over the others – reward-based question

We categorised participants' responses about why they selected a particular HWR as the one with the best performance for the reward-based question. Each response was associated with one or two of the following seven themes: *number of errors, generic, others' opinion, type of errors, animation, speed,* and *order*. Figure 31 illustrates the frequencies of these themes. The themes were the same as those emerged from previous studies, with the exception of *order* which was used for the following comment: "I didn't find any errors in the first program, and it is the first on the list."

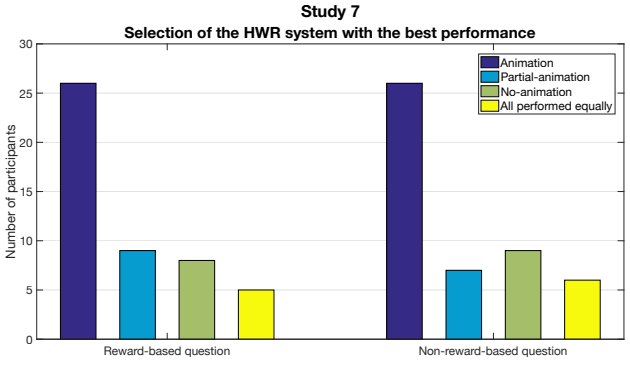

Figure 30. Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 7. Number of participants on the y-axis.

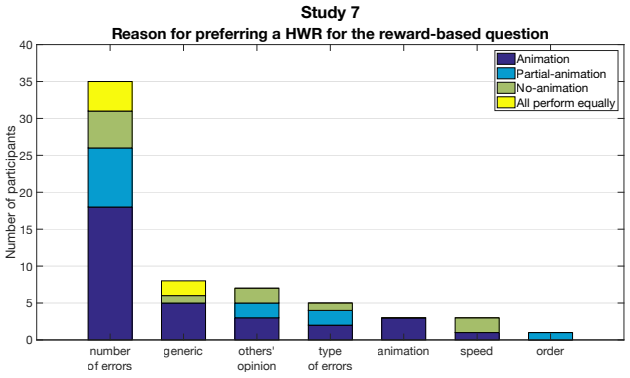

Figure 31. Reasons expressed by participants for preferring a HWR in the *animation*, *partial-animation* or *no-animation* condition for the reward-based question. Number of participants on the y-axis.

9.2.3. Reasons for choosing one HWR system over the others – non-reward-based question

The comments of the participants about why they selected a particular HWR as the one with the best performance for the non-reward-based question were categorised through thematic analysis. Each response was associated with one or two themes, with seven themes used in total: *number of errors, type of errors, generic, speed, animation, others' opinion,* and *random*. Figure 32 illustrates the frequencies of these themes. While the first six themes are the same as above. The new theme *random* was associated with the response "I am not sure, all of them seemed to perform similarly, but it looks like the third was maybe the best? Not sure."

For the majority of participants, 25 out of 48, their answers were the same (in terms of themes) for the reward-based question. Only 23 participants answered this question differently than the previous one, and of these 23 only 4 changed their selection. In more detail, the first participant who selected a different system commented that he believed that other participants would choose the system in the *partial-animation* condition because it was the first system they saw and they would remember ("I think people might choose the first one because their attention and focus will be greater when viewing the first program compared to the second and third."). The second participant considered the majority would select that the *three systems* have the same performance for the reason that had the same performance. However, he felt that the system in the *no-animation* condition was the one with the best performance because it was the only one that produced all of its output in one go. The third participant considered that other participants would be distracted to evaluate the system in the *animation* condition ("People will lose some concentration, so they will not be accurate.") Finally, the last participant changed their selection from *three systems* performed equally to the system in the *animation* condition. Regarding 19 participants who provided different reasons without changing their selection, 4 participants changed their answers from "number of errors' to "generic". Additionally, 3 participants' answers changed from "number of errors" to "type of errors", 2 participants' answer changed from "generic" to "generic and type of errors", 2 participants' answers change from "generic" to "speed". Moreover, three participants' answers changed from "others' opinion" to "random" to "number of errors" to "type of errors". In another case, two participants changed their answers from "type of errors" to "number of errors" to "speed". In another case the answer was changed from "speed" to "generic", and on another, the answer changed from "generic" to "number of errors". Finally, in the last case, the participant changed the answer from "number of errors and generic" to just "number of errors".

For this study, five participants changed their choice of which HWR system performed the best. The two participants that changed their choice from *partial-animation* and three programs to *animation* condition mentioned that they considered that the *animation* condition has the best performance and so other participants will pay more attention to the system in that condition. The two participants that changed their choice from *animation* and all performed equally to *no-animation* commented that they considered that this condition is faster and this will have motivated people to select the other conditions. Finally, the participant that changed from *animation* condition to all performed equally considered that most of the participants will select the animation because this will distract them.

9.2.4. Performance ratings

A CHI-squared test revealed statistically significant differences in the performance ratings, $\chi^2(2) = 9.73, p < 0.01$. Post-hoc analysis with pairwise Wilcoxon Signed-ranks tests with significance level set at $p < 0.05$, revealed Median for the performance evaluation for the *animation*, *partial-animation*, and *no-animation* were $Mdn = 4.5$, $Mdn = 4$, and $Mdn = 4$, respectively. There were significant differences between the *animation* and the *no-animation* conditions ($Z = 3.037, p < 0.01, r = 0.31$), and also between the *partial-animation* and the *no-animation* conditions ($Z = 2.64, p < 0.05, r = 0.25$). No significant differences were instead found between the *animation* and *partial-animation* conditions ($Z = 0.89, p > 0.05, r = 0.09$). Figure 33 shows participants evaluation of the performance of the systems.

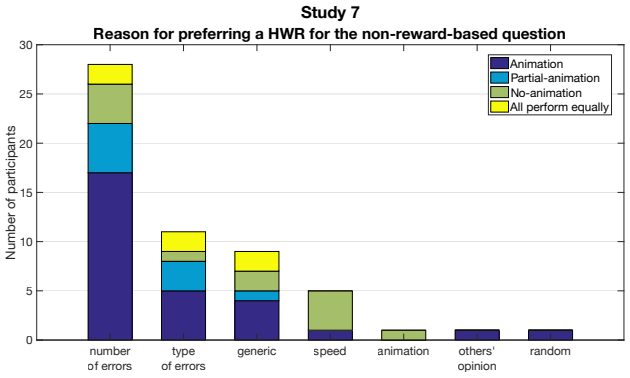

Figure 32. Reasons expressed by participants for preferring a HWR in the *animation*, *partial-animation*, or *no-animation* condition for the non-reward-based question. Number of participants on the y-axis.

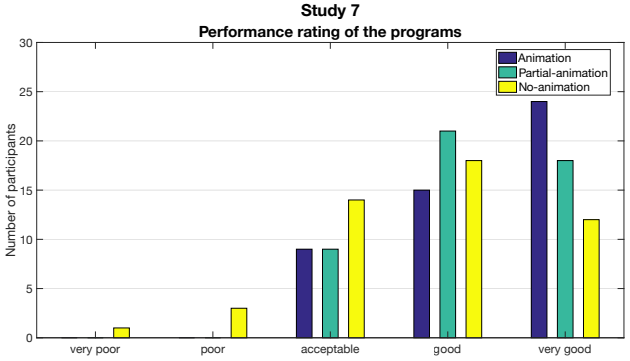

Figure 33. Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions. Number of participants on the y-axis.

9.2.5. Discussion

The results of Study 7 suggest that *any amount* of animation seems to influence users' perception of the performance of the system: statistically significant differences in the Likert-scale ratings were found both between *no-animation* and *animation* and between *no-animation* and *partial-animation*, while no statistically significant differences were found between *animation* and *partial-animation*. However, in terms of choosing the system with the best performance, the majority of participants opted for the *animation* condition, rather than any of the other 3 options, regardless of the experimental reward. Indeed, almost three times as many participants opted for the system in the *animation* condition compared to the one in the *partial-animation* one. In contrast to the Likert-scale results, the selection results suggest that the amount of animation does play some role in users' perception of performance. So perhaps the lack of statistical significance mentioned above could be a limitation of our sample size.

Once again, similar to prior studies the qualitative data from Study 7 indicates that participants took the task seriously and engaged with it.

10. Study 8

In all studies reported so far, the presence of animation was the only difference across the systems our participants evaluated. The performance of the various systems, defined in terms of number of errors produced by the system was kept constant. To further characterise the effect we identified, we decided to test what level of imbalance in the performance level of the system being compared would "break the illusion" created by the animation. In other words: how many additional errors can the animation cover? Study 8 was designed to address this question, by comparing pairs of systems with different numbers of mistakes.

10.1. *Method*

10.1.1. Study Design

The study design was based on the design of Study 2, but we additionally divided participants into two groups, each group corresponding to a different number of errors: *9-errors* group and *10-errors* group. For each group the experiment was identical to Study 2, with the exception that the *animation* condition the number of errors indicated by the group name. The *no-animation* condition always included just 8 errors, as it was in Study 2 (in earlier studies the number of errors was the same across the conditions).

The reward structure and amounts were identical to Study 2, with a fixed amount of $1.17 being paid to all participants, plus a bonus of the same amount for those who answer the *reward-based* question in the majority.

10.1.2. Participants

Participants were recruited through MTurk, with the same two restrictions as in Study 2. The sample size was 32, double than what it was for Study 2, to account for the split of participants into 2 groups. The age ranged from 19 to 62 ($M = 27, SD = 6.80$), 22 males (69%) and 10 females (31%). All except for 3 of the participants reported to be

United States nationals, the remaining ones being Algerian, Canadian and Japanese. The education levels of the participants ranged from secondary school level to masters' degree level or equivalent. Overall, 6 participants had a masters' degree, 20 had a university degree, and 6 completed secondary school. One of the participants reported knowing Filipino.

10.1.3. Equipment

The same Web application used for Studies 1 and 2 was used for Study 8, with the only difference of the number of errors in the *animation* condition, as described above.

10.1.4. Procedure

The procedure was the same as Study 2.

10.2. *Results*

10.2.1. Selection of the system with the best performance

9-errors group. For the *reward-based* question, 9 of the 16 participants (56%) chose the system in the *animation* condition as the one with the best performance. Additionally, 7 participants (25%) selected the system in the *no-animation* condition as the one with the best performance, while the remaining 3 (19%) indicated that both systems had the same performance level. Only 1 participant answered the *non-reward-based* question differently than the *reward-based* one, changing the answer from "animation" to "both systems". These results are illustrated in Figure 34.

10-errors group. For the *reward-based* question, 4 of the 16 participants (25%) chose the system in the *animation* condition as the one with the best performance, 6 participants (37.5%) selected the system in the *no-animation* condition as the one with the best performance, while the remaining 6 (37.5%) indicated that both systems had the same performance level. Only four participants answered the *non-reward-based*

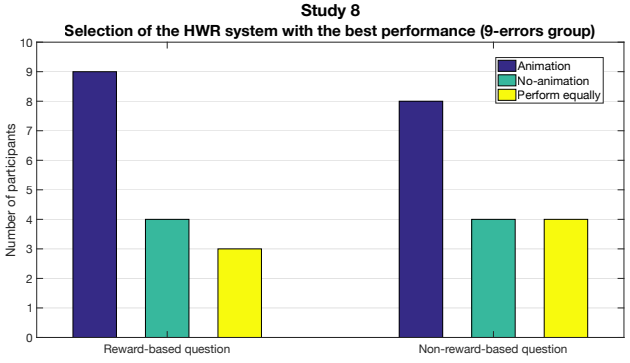

Figure 34. Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 8 for the 9-errors group. Number of participants on the y-axis.

question differently than the *reward-based* question. The first participant changed from "both systems" to "animation", the second participant changed from "both systems" to "no-animation". The third participant changed from "animation" to "both systems". Finally, the last participant change from "no-animation" to "both systems". With these changes the amount of participant that select one system in for the *reward-based* question was similar for the *non-reward-based* question. These results are illustrated in Figure 35. The participant who reported knowing Filipino was in this group, and she selected the system in the *animation* condition as the one having the best performance for the *reward-based* question, while she indicated that both programs had the same performance for the *non-reward-based* question.

10.2.2. Reasons for choosing one system over the other – reward-based question.

We categorised participants' comments into themes based on the reasons why they chose a particular HWR as the one majority of participant will choose as the one with the best performance. Each response was associated with one or two themes, with six themes found in total: *number of errors, generic, animation, type of errors,* and *others' opinion*. The five themes are the same as in Study 4. Figure 36 illustrates the frequencies of these themes, also classified on the reward-based question (*animation* or *no-animation*) for the **9-errors group**, while Figure 37 illustrates the frequencies for the **10-errors group**. The participant who reported knowing Filipino was in the 10-errors group, and her answer was associated with the theme *type of errors*.

10.2.3. Reasons for choosing one system over the other – non-reward-based question.

We grouped participants' responses into themes based on the reasons why they selected a particular HWR as the one with the best performance. Each response was associated with one or two themes, with four themes found in total: *number of errors, generic, speed* and *type of errors*. The themes categorised participants' comments as we did in the previous studies. Figure 38 illustrates the frequencies of these themes, also

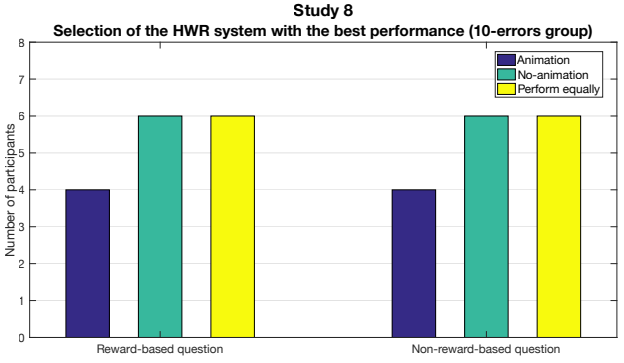

Figure 35. Selection of the participants for preferring a HWR for the reward-based and non-reward-based questions in Study 8 for the 10-errors group. Number of participants on the y-axis.

classified on the individual preference (*animation* or *no-animation*) for the **9-errors group**, while Figure 39 illustrates the frequencies for the **10-errors group**. Similar to the previous subsection, the answer of the participant who reported knowing Filipino was associated with the theme *type of errors*.

Only one participant in the group 9-errors changed her selection from *animation* condition to "both systems" performed equally. The reason behind why the participant changed her choice is due to the fact the participant could see how the system worked in its task.

For the purpose of the analysis we report each group corresponding to a number of errors separately. This applies both to the quantitative and the qualitative analysis.

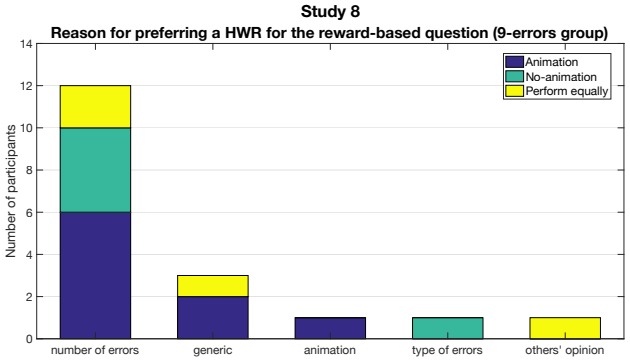

Figure 36. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question for the 9-errors group. Number of participants on the y-axis.

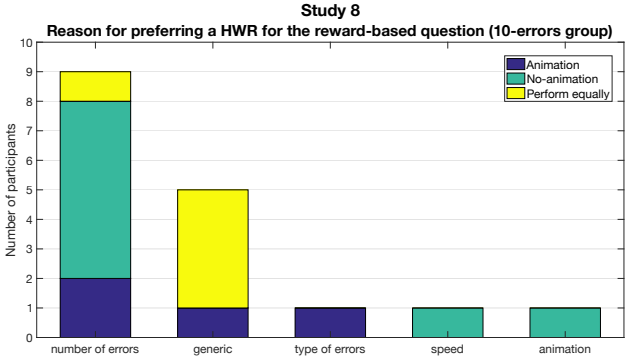

Figure 37. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the reward-based question for the 10-errors group. Number of participants on the y-axis.

10.2.4. Performance ratings

9-errors group. A Wilcoxon Signed-ranks Test revealed that the performance evaluation was higher in the *animation* condition ($Mdn = 4$) than in the *no-animation* condition ($Mdn = 4$), ($Z = 2.183, p < .05, r = 0.39$). Figure 40 shows participants evaluation of the performance of the systems.

10-errors group. A Wilcoxon Signed-ranks test did not reveal statistical significance suggesting the positive effect of animation was cancelled ($Z = 0, p = 1, r = 0$). Figure 41 shows participants evaluation of the performance of the systems.

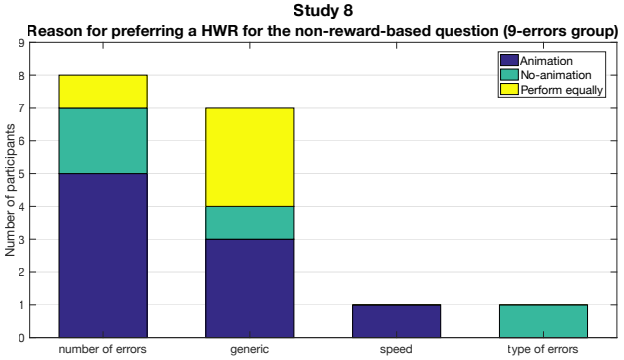

Figure 38. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the non-reward-based question for the 9-errors group. Number of participants on the y-axis.

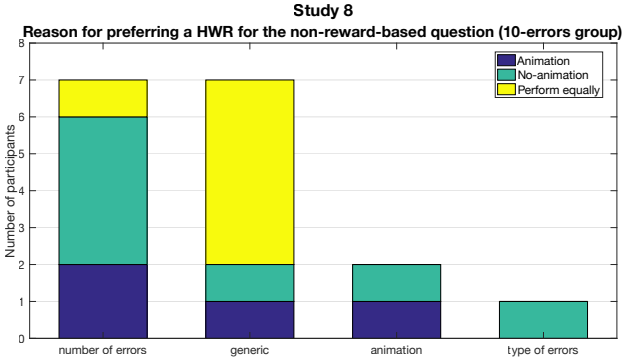

Figure 39. Reasons expressed by participants for preferring a HWR in the *animation* or *no-animation* condition for the non-reward-based question for the 10-errors group. Number of participants on the y-axis.

10.3. *Discussion*

The results of Study 8 indicate that the effect of animation cues on participants' perception of the system performance holds, to some extent, even when comparing two systems that have different performance levels. In particular, within the 9-errors group most participants selected the system in the *animation* condition, even when it produced one additional mistake compared to the system in the *no-animation* condition (corresponding to a performance degradation of 12.5%). When the difference in number of errors produced by the two systems becomes 2 (the 10-errors group, which corresponds to a performance degradation of 25%), the animation system is no longer selected as the one with the best performance by the majority of participants, but only by 4 participants (25%). However, even in the 10-errors group 6 participants (37.5%)

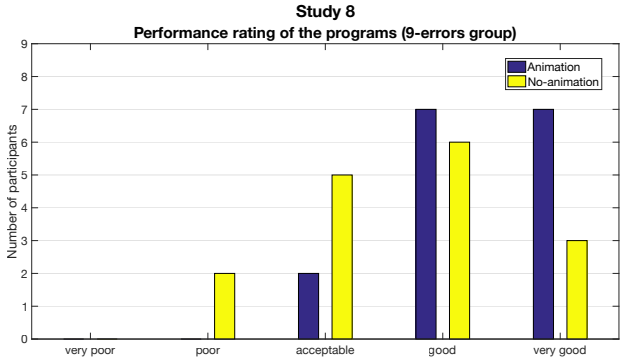

Figure 40. Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions for the 9-errors group. Number of participants on the y-axis.

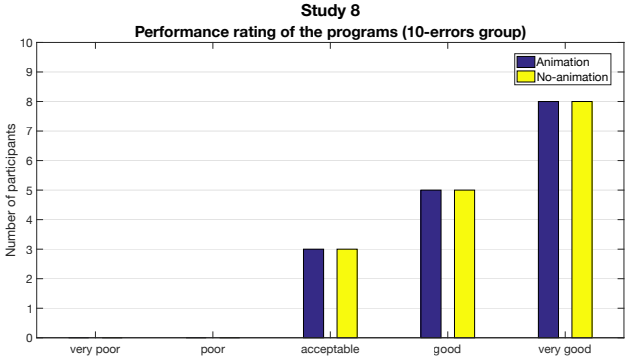

Figure 41. Likert-scale of participants' evaluation of the performance of the systems for the *animation* and *no-animation* conditions for the 10-errors group. Number of participants on the y-axis.

suggested that the two systems have the same performance, and that's as many as those who correctly selected the system in the *no-animation* condition as the one with the best performance. This finding is reinforced by the qualitative data, this shows that in both the 9-errors group and the 10-errors group, some participants suggested that there are fewer errors in the *animation* condition compared to the *no-animation* condition.

More in general, from this study, we can learn that the positive effect of animation cues can persist even when a system's performance is degraded. In contrast, Kim & Hinds (2006) showed that people who worked in a cooperative environment, blame others when a robot has a problem, and this one delivers an explanation of its failure. In more detail, people did not consider that the problem was a malfunction of the robot. Instead, they considered that the problems were due to interactions with other people. In contrast, our findings show that the *animation* cues tend to hide a possible malfunction of the system.

Finally, the answers submitted by the participant who reported knowing the Filipino language suggest once again that the knowledge of the language did not influence the behaviour in our study.

11. Summary

Our initial three studies revealed that animation cues integrated into the GUI of a smart system can affect people's perception of the system performance, extending and generalising the results reported by Garcia et al. (2016) for physical motion cues and vacuum cleaning robots. In particular, in Study 1 (N=16) participants reported a HWR system to perform better when animation cues are displayed than when they are not. Study 2 (N=16) replicated the same experiment on MTurk, extending the initial results to a less controlled environment, and demonstrating that further studies could be conducted on the online platform. Study 3 (N=16) demonstrates that the effect of animation is not specific to the type of smart system used in Studies 1 and 2, similar results were observed also for a POS tagging system, one which involves a type of data processing that is less inherently visual than the HWR system.

Studies 4, 5 and 6 were designed to look for an explanation for this effect. In particular, Study 4 (N=16) examined and ruled out the possibility that the animation cues may induce users to perceive that the system recognises the handwriting as a person would, and so appears to be "as smart as a person." Study 5 (N=16) provides an initial exploration of the relationship between the animation and participants' mental model of the system. Study 6 (N=64) probed such relationship further: its results suggest that animation cues affect participants' perception of the system performance, only if the animation matches their mental model of the system. More in general, the results of Studies 5 and 6 suggest that if the animation cues are largely compatible with a user's mental model of the system, then they act as a *reminder* for how the system works and they can increase the user's confidence in the system (at least compared to alternative systems for which they have no cues about how it works).

Once we found the reason behind why *animation* cues influence participants' perception of smart systems, we designed and conducted two further studies to characterize the observed phenomenon more in detail. In Study 7 (N=48) We analysed whether the amount of animation shown would influence participants' perception of a system's performance. The results indicate that any amount of animation seem to have potential to influence users perception of a system performance. Finally, Study 8 (N=32)

assessed the effect of *animation* cues when a system's performance actually decreases, compared to the alternative system where no animation is integrated. The results of Study 8 show that that the effect of animation cues on participants' perception of the system performance holds, to some extent, even in this case. In particular, most participants still favour the system with animation even when it makes one error more than the the system with no animation. However, when the number of extra errors becomes 2, the effect decreases.

12. Implications

Our studies bear strong implications for the design of user interfaces for smart systems, and in particular the design of visual feedback around such systems. Examples of these systems include, among many others, mobile and web applications like translators, recommender systems or even chatbots in e.g. automated customer support. Overall, our results imply that designers should be aware that including animations in the UI can bias users' perception of how well a smart system works. This could also affect mundane animations such as loading screens, transition animations or even decorative animations. Our findings point to the need of more investigation in this area. In particular, we found that if the animations are largely compatible with users mental model of the system, they can lead users to have a more positive perception of the system performance. So, if designers intend to try to deliberately induce such bias to give users a more favourable impression of their system, particular care then needs to be taken to make sure that users mental model can be reliably predicted, so that the animation can be made compatible with it.

However, it is perhaps even more important to be aware of the risk that animations may inadvertently lead users to rely on the results of a smart system more than they should. Indeed our last study indicates that animations can even 'cover up' some of the errors made by the system. Such over-reliance can have undesirable consequences, if not even dramatic, especially for safety-critical applications (Parasuraman & Riley (1997)). Additional research is needed to understand the potential consequences.

12.1. *Further Research Opportunities*

While our results show an evident effect, they also open up a number of new research questions, for example: does it apply to other screen-based systems or maybe even any application? Reflecting on the physical motion cues presented by Garcia et al. (2016), it would be interesting to explore if people also rated the robot's performance higher because seeing the motion made them think that they understand how the robot works.

13. Conclusion

In this paper, we presented eight studies designed to explore whether visual animation cues can change people's perception in the evaluation of smart systems and what characteristics the animations need to have. Indeed, our results suggest that displaying a high detail of animations that match people's mental model as animation cues can influence people's perception of the performance of smart screen-based systems even these systems have a minimal decrease in their performance.

In general, our results indicate that this modality has potential to improve user ratings, as the display of animation can change how people perceive and evaluate the system's performance. We expect that the results presented in this paper will stimulate designers to integrate animations as a feedback of their systems, and researchers to explore this area further.

References

Ariely, D. (2008). *Predictably irrational*. HarperCollins New York.

Bakhshi, S., Shamma, D. A., Kennedy, L., Song, Y., de Juan, P., & Kaye, J. J. (2016). Fast, cheap, and good: Why animated gifs engage us. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 575–586). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2858036.2858532 doi:

Balijepally, V., Nerur, S., & Mahapatra, R. (2015). Task mental model and software developers' performance: An experimental investigation. *Communications of the Association for Information Systems*, *36*(1), 4.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. Retrieved from http://www.tandfonline.com/doi/abs/10.1191/1478088706qp063oa doi:

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical turk. *Perspectives on Psychological Science*, *6*(1), 3-5. Retrieved from http://dx.doi.org/10.1177/1745691610393980 (PMID: 26162106) doi:

Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Chuang, J., Manning, C. D., & Heer, J. (2012, October). Without the clutter of unimportant words: Descriptive keyphrases for text visualization. *ACM Trans. Comput.-Hum. Interact.*, *19*(3), 19:1–19:29. Retrieved from http://doi.acm.org/10.1145/2362364.2362367 doi:

Detenber, B. H., & Reeves, B. (1996). A bio-informational theory of emotion: Motion and image size effects on viewers. *Journal of Communication*, *46*(3), 66–84. Retrieved from http://dx.doi.org/10.1111/j.1460-2466.1996.tb01489.x doi:

Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., & Cudré-Mauroux, P. (2015). The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th international conference on world wide web* (pp. 238–247). Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. Retrieved from http://dl.acm.org/citation.cfm?id=2736277.2741685

Dittrich, W. H., & Lea, S. E. G. (1994). Visual perception of intentional motion. *Perception*, *23*(3), 253-268. Retrieved from http://pec.sagepub.com/content/23/3/253.abstract doi:

Fritz Heider, M. S. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243-259. Retrieved from http://www.jstor.org/stable/1416950

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, *21*(12), 1845-1853. Retrieved from http://pss.sagepub.com/content/21/12/1845.abstract doi:

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 669. doi:

Garcia, P. G., Costanza, E., Ramchurn, S. D., & Verame, J. K. M. (2016). The potential of physical motion cues: Changing people's perception of robots' performance. In *Proceedings of the 2016 acm international joint conference on pervasive and ubiquitous computing* (pp. 510–518). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2971648.2971697 doi:

Gelman, R., Durgin, F., & Kaufman, L. (1995). Distinguishing between animates and inanimates: Not by motion alone. *Causal cognition: A multidisciplinary debate*, 150–184.

Germine, L., Nakayama, K., Duchaine, B., Chabris, C., Chatterjee, G., & Wilmer, J. (2012). Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, *19*(5), 847-857. Retrieved from http://dx.doi.org/10.3758/s13423-012-0296-9 doi:

Kazai, G., Kamps, J., & Milic-Frayling, N. (2013). An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, *16*(2), 138-178. Retrieved from http://dx.doi.org/10.1007/s10791-012-9205-0 doi:

Kieras, D. E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, *8*(3), 255–273. Retrieved from http://dx.doi.org/10.1207/s15516709cog0803 doi:

Kim, T., & Hinds, P. (2006, Sept). Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction. In *Robot and human interactive communication, 2006. roman 2006. the 15th ieee international symposium on* (p. 80-85). doi:

Krause, J., Perer, A., & Ng, K. (2016). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 5686–5697). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2858036.2858529 doi:

Lim, B. Y., & Dey, A. K. (2011). Design of an intelligible mobile context-aware application. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services* (pp. 157–166). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2037373.2037399 doi:

Lim, B. Y., Dey, A. K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 2119–2128). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1518701.1519023 doi:

Lyons, J. (2013). Being transparent about transparency: A model for human-robot interaction.. Retrieved from http://www.aaai.org/ocs/index.php/SSS/SSS13/paper/view/5712

Mason, W., & Watts, D. J. (2010, May). Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.*, *11*(2), 100–108. Retrieved from http://doi.acm.org/10.1145/1809400.1809422 doi:

Michotte, A. (1963). The perception of causality.

Norman, D. A. (2013). *The design of everyday things: Revised and expanded edition.* Basic books.

Nowacka, D., Hammerla, N. Y., Elsden, C., Plötz, T., & Kirk, D. (2015). Diri - the actuated helium balloon: A study of autonomous behaviour in interfaces. In *Proceedings of the 2015 acm international joint conference on pervasive and ubiquitous computing* (pp. 349–360). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2750858.2805825 doi:

O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., & Höllerer, T. (2008). Peerchooser: Visual interactive recommendation. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1085–1088). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1357054.1357222 doi:

Pantelis, P. C., & Feldman, J. (2012). Exploring the mental space of autonomous intentional agents. *Attention, Perception, & Psychophysics*, *74*(1), 239–249.

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

Park, D., & Lee, J.-H. (2010a). Investigating the affective quality of motion in user interfaces to improve user experience. In H. S. Yang, R. Malaka, J. Hoshino, & J. H. Han (Eds.), *Entertainment computing - icec 2010: 9th international conference, icec 2010, seoul, korea, september 8-11, 2010. proceedings* (pp. 67–78). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-15399-0 doi:

Park, D., & Lee, J.-H. (2010b, December). Understanding how the affective quality of motion is perceived in the user interface. *Comput. Entertain.*, *8*(2), 14:1–14:11. Retrieved from

http://doi.acm.org/10.1145/1899687.1899696 doi:

Popović, J., Seitz, S. M., & Erdmann, M. (2003, October). Motion sketching for control of rigid-body simulations. *ACM Trans. Graph.*, *22*(4), 1034–1054. Retrieved from http://doi.acm.org/10.1145/944020.944025 doi:

Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press Cambridge, UK.

Schlottmann, A., & Surian, L. (1999). Do 9-month-olds perceive causation-at-a-distance? *Perception*, *28*(9), 1105-1113. Retrieved from http://pec.sagepub.com/content/28/9/1105.abstract doi:

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of personality and social psychology*, *45*(3), 513.

Talbot, J., Lee, B., Kapoor, A., & Tan, D. S. (2009). Ensemblematrix: Interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1283–1292). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1518701.1518895 doi:

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*(8), 943-951. Retrieved from http://pec.sagepub.com/content/29/8/943.abstract doi:

Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007). How it works: A field study of non-technical users interacting with an intelligent system. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 31–40). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1240624.1240630 doi:

Tversky, A., & Kahneman, D. (1985). The framing of decisions and the psychology of choice. In *Environmental impact assessment, technology assessment, and risk analysis* (pp. 107–129). Springer.

Verame, J. K. M., Costanza, E., & Ramchurn, S. D. (2016). The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study. In *Proceedings of the 2016 chi conference on human factors in computing systems* (pp. 4908–4920). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2858036.2858369 doi:

Vermeulen, J. (2010). Improving intelligibility and control in ubicomp. In *Proceedings of the 12th acm international conference adjunct papers on ubiquitous computing - adjunct* (pp. 485–488). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1864431.1864493 doi:

Ware, C. (2012). *Information visualization: perception for design*. Elsevier.

Yang, R., & Newman, M. W. (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* (pp. 93–102). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2493432.2493489 doi:

Yatani, K., Novati, M., Trusty, A., & Truong, K. N. (2011). Review spotlight: A user interface for summarizing user-generated reviews using adjective-noun word pairs. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 1541–1550). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1978942.1979167 doi:

Appendix C

Ethics Forms

In the following pages, we include the ethics form we submit to the FPSE Ethics Committee at the University of Southampton for the Studies 1, 2, 3, and 4 we ran in the University's lab. Additionally, we add the consent form that our participants signed before they took part in the same studies.

FPSE Ethics Committee
FPSE EC Application Form Ver 6.6d

Refer to the *Instructions* and to the *Guide* documents for a glossary of the key phrases in **bold** and for an explanation of the information required in each section. The *Templates* document provides some text that may be helpful in presenting some of the required information.

Replace the highlighted text with the appropriate information.

Note that the size of the text entry boxes provided on this form does **not** indicate the expected amount of information; instead, refer to the *Instructions* and to the *Guide* documents in providing the complete information required in each section. Do **not** duplicate information from one text box to another.

Reference number: **ERGO/**FPSE**/17155**	Version: 1	Date: 2015-08-10
Name of **investigator**(s): Pedro Garcia Garcia		
Name of supervisor(s) (if student **investigator**(s)): Enrico Costanza Sarvapali Ramchurn		
Title of study: Visualisation of automation, as a tool for changing people's perception		
Expected study start date: 08/10/2015	Expected study end date: 09/25/2016	
Note that the dates requested on the "IRGA" form refer to the start and end of *data collection*. These are not the same as the start and end dates of the study for which approval is sought. Note that approval must be obtained before the study commences; retrospective approval cannot be given.		

The investigator(s) undertake to:

- Ensure the study Reference number ERGO/FPSE/17155 is prominently displayed on all advertising and study materials, and is reported on all media and in all publications;
- Conduct the study in accordance with the information provided in the application, its appendices, and any other documents submitted;
- Conduct the study in accordance with University policy governing research involving human **participants** (http://www.southampton.ac.uk/ris/policies/ethics.html);
- Conduct the study in accordance with University policy on data retention (http://www.southampton.ac.uk/library/research/researchdata/);
- Submit the study for re-review (as an amendment through ERGO) or seek FPSE EC advice if any changes, circumstances, or outcomes materially affect the study or the information given;
- Promptly advise an appropriate authority (Research Governance Office) of any adverse study outcomes, changes, or circumstances (via an adverse event notification through ERGO);
- Submit an end-of-study form as may be required by the Research Governance Office upon completion of the study.

REFER TO THE INSTRUCTIONS DOCUMENT WHEN COMPLETING THIS FORM.

PRE-STUDY

Characterise the proposed **participants**
The participants would be gathered from the University of Southampton (e.g. under graduate students, PhD students, staff and researchers).

Describe how **participants** will be approached
- Participants will be approached through mailing lists of selected research groups (e.g. our research groups), and through personal contacts - Posters will be affixed on campus If any non-FPSE e-mail lists are used, justify their use

Describe how inclusion and/or exclusion criteria will be applied (if any)
- Because one tasks involve reading a considerable amount of text, people affected by dyslexia will be excluded from the study. This will be based on self-reporting. When participants express interest in the study, before scheduling with them a time to take part, we will let them know that one tasks involve reading a considerable amount of text, so this may not be appropriate for someone with dyslexia, so if they want to go ahead, they should please confirm this is fine by them and they do not have dyslexia.

Describe how **participants** will decide whether to take part
- Participants will be provided the Participant Information via email upon acceptance of invitation to take part. - Those who have accepted will be given a link to an online scheduling system (Doodle.com) to reserve a slot. Once everyone has done this, they will be e-mailed containing details of when and where their allocation for the study will be. - Participants will be asked to sign the consent form before the study commences.

Participant Information

Provide the **Participant Information** in the form that it will be given to **participants** as an appendix. All studies must provide **participant information**.

Consent Form

Provide the **Consent Form** (or the request for consent) in the form that it will be given to **participants** as an appendix. All studies must obtain **participant** consent. Some studies may obtain verbal consent, other studies will require written consent, as explained in the *Instructions* and *Guide* documents.

DURING THE STUDY

Describe the study procedures as they will be experienced by the **participant**
In the experiment, participants will be asked to engage in two experiments. First experiment is to observe a robot's outcome and evaluate robot's performance. Second experiment is computer based. Participants will do one experiment at the time. **First experiment: Visualise and evaluate robot's outcome** The first experiment involves to observe and evaluate the rooms where two robots are going to work before they start their task. This evaluation is going to be in the questionnaire that they are going to receive. Meanwhile robots work in their task, participants are going to wait in a second room. Once robots finish their task, participants will return to robot's rooms and evaluate robots' performance in the questionnaire that they have. At the end of the study, participants who had chosen the robot with the best performance, they are going to receive a reward of £5. This mechanism of reward is with the aim to enhance people to evaluate more consciously robot's performance. **Second Experiment: Checking and correcting automatic handwriting recognition.** For this second experiment, participants are going to check and correct the results of a semi-automatic system that recognizes handwriting (this is actually simulated). Participants are going to check three documents from three different systems and they need to find mistakes that the systems made on the documents. After this, they will evaluate which system has a better performance according to their experience working in finding mistakes on the documents. At the end of the study, participants who had chosen the robot with the best performance, they are going to receive a reward of £5. This mechanism of reward is with the aim to enhance people to evaluate more consciously robot's performance. **Reward Mechanism** After we finish the study, participants for the first experiment who choose the robot with the best performance are going to be contact by email in order to receive their reward of £5. The same case is going to be for the second experiment. The maximum reward possible is £10 if the participants are involving in the two experiments. Before the study begins, the participants will be given the consent form to read and sign. The participants will then be introduced to the experiments. This should take around 5 minutes. The first experiment is going to last for about 10 minutes and the second experiment will be last 15 minutes. After the experiments, participants are going to receive a questionnaire where they are going to evaluate the performance of the systems and their perception of them.

Identify how, when, where, and what kind of data will be recorded (not just the formal research data, but including all other study data such as e-mail addresses and signed consent forms)
- Upon completion of the consent form, this will be collected (prior to the beginning of the study).

- With the questionnaire that participants will answer after they finish with the experiments.
- No sensitive data will be recorded.

Participant questionnaire

As an appendix, if using a questionnaire, reproduce any and all **participant** questionnaires or data gathering instruments in the exact forms that they will be given to or experienced by **participants**. If conducting less formal data collection, provide specific information concerning the methods that will be used to obtain the required data.

POST-STUDY

Identify how, when, and where data will be stored, processed, and destroyed

Data file will be stored in a password-protected computer within the University of Southampton. Any physical data shall be destroyed once the study finish on 20th September 2016, but its content will be moved to the password-protected computer and the information will be deleted at the end of my PhD on 25th September 2018.

- Any information published from this study will be done in aggregate or anonymous form.

If Study Characteristic M.1 applies, provide this information in the **DPA Plan** as an appendix instead and do not provide explanation or information on this matter here.

STUDY CHARACTERISTICS

(L.1) The study is funded by a commercial organisation: **No**

If 'Yes', provide details of the funder or funding agency here

(L.2) There are **restrictions** upon the study: **Yes**

If 'Yes', explain the nature and necessity of the **restrictions** here

The second experiment requires comprehensive reading and grammar correction and thus we require people with no dyslexia to make sure that the results will be comparable, and to avoid distress to people suffering dyslexia.

(L.3) Access to **participants** is through a third party: **No**

If 'Yes', provide evidence of your permission to contact them as a separate appendix. Do not provide explanation or information on this matter here

(M.1) **Personal data** is collected or processed: **Yes**
Data will be processed outside the UK: **No**

If 'Yes' to either question, provide the **DPA Plan** as a separate appendix. Do not provide information or explanation on this matter here. Note that using or retaining e-mail addresses, signed consent forms, or similar study-related **personal data** requires M.1 to be "Yes"

(M.2) There is **inducement** to **participants**: **Yes**

If 'Yes', explain the nature and necessity of the inducement here

The inducement is used to attract people to join the study and as an incentive for spending their time.

(M.3) The study is **intrusive**: **No**

If 'Yes', provide the **Risk Management Plan** and the **Debrief Plan** as appendices, and explain here the nature and necessity of the intrusion(s)

(M.4) There is **risk of harm** during the study: **No**

If 'Yes', provide the **Risk Management Plan**, the **Contact Information**, and the **Debrief Plan** as appendices, and explain here the necessity of the risks

(M.5) The true purpose of the study will be hidden from **participants**: **No** (doho)
The study involves **deception** of **participants**: **No**

If 'Yes' to either question, provide the **Debrief Plan** as an appendix, and explain here the necessity of the deception

(M.6) **Participants** may be minors or otherwise have **diminished capacity**: **No** (doho)

If 'Yes', AND if one or more Study Characteristics in categories M or H applies, provide the **Risk Management Plan** and the **Contact Information**, as appendices, and explain here the special arrangements that will be put in place that will ensure informed consent

(M.7) **Sensitive data** is collected or processed: **No**

If 'Yes', provide the **DPA Plan** as a separate appendix. Do not provide explanation or information on this matter here

(H.1) The study involves: **invasive** equipment, material(s), or process(es); or **participants** who are not able to withdraw at any time and for any reason; or animals; or human tissue; or biological samples: **No**

If 'Yes', provide further details and justifications as one or more separate appendices. Do not provide explanation or information on these matters here. Note that the study will require separate approval by the Research Governance Office

Technical details

If one or more Study Characteristics in categories M.3 to M.7 or H applies, provide the description of the technical details of the experimental or study design, the power calculation(s) which yield the required sample size(s), and how the data will be analysed, as separate appendices. Do not provide explanation or information on these matters here.

APPENDICES (AS REQUIRED)

While it is preferred that this information is included here in the Study Protocol document, it may be provided as separate documents.

If provided separately, be sure to name the files precisely as "Participant Information", "Questionnaire", "Consent Form", "DPA Plan", "Permission to contact", "Risk Management Plan", "Debrief Plan", "Contact Information", and/or "Technical details" as appropriate.

If provided separately, each document must specify the reference number in the form ERGO/FPSE/xxxx, its version number, and its date of last edit.

Appendix (i): **Participant Information** in the form that it will be given to **participants.**

Appendix (ii): Data collection plan / Questionnaire in the form that it will be given to **participants.**

Appendix (iii): **Consent Form** in the form that it will be given to **participants.**

Appendix (iv): **DPA Plan.**

Appendix (v): Evidence of permission to contact **participants** or prospective **participants** through any third party.

Appendix (vi): **Risk Management Plan**.

Appendix (vii): **Debrief Plan**.

Appendix (viii): **Contact Information**.

Appendix (ix): Technical details of the experimental or study design, the power calculation(s) for the required sample size(s), and how the data will be analysed.

Appendix (x): Further details and justifications in the case of **invasive** equipment, material(s), or process(es); participants who are not able to withdraw at any time and for any reason; animals; human tissue; or biological samples.

Appendix (i) Participant Information template

Participant Information

Ethics reference number: **ERGO/**FPSE**/17155**	Version: 1	Date: 2015-08-10
Study Title: Visualisation of automation, as a tool for changing people's perception		
Investigator: Pedro Garcia Garcia, Enrico Costanza, Sarvapali Ramchurn		

Please read this information carefully before deciding to take part in this research. If you are happy to participate you will be asked to sign a consent form. Your participation is completely voluntary.

What is the research about? This is a research project, which will evaluate how people evaluate system's performance.

Why have I been chosen? You have been approached because you expressed an interest in participating.

What will happen to me if I take part? You will participate in two experiments.

First experiment: visualise and evaluate robot's outcome

In the first experiment, you would evaluate the rooms where the robots are going to work on them. This evaluation you have to make it in the questionnaire that we are going to give you at the beginning of the study. After the evaluation you will be waiting in a second room until robots finish their task. Finally, you will return to the room where robots worked and you will evaluate their performance. Make sure to choose which robot you think that the majority are going to select that has the best performance, because according to this we will contact you at the end of the study to give you the reward of £10 in the case that you choose the robot that the majority selected.

Second experiment: Checking and correcting automatic handwriting recognition

In the second experiment, you will check and find mistakes in the results of a semi-automatic system that recognize handwriting. You are going to check three documents from three different systems, then you will evaluate which system has a better performance according to your experience working in finding mistakes on the three documents. Make sure to choose which robot you think that the majority are going to select that has the best performance, because according to this we will contact you at the end of the study to give you the reward of £10 in the case that you choose the robot that the majority selected.

The experiments are going to be in two different days. In overall, your participation will be around 25 minutes for the two experiments. You can have the opportunity to win £20 maximum reward for your participation.

Are there any benefits in my taking part? It is expected that you participate in two experiments, this that can give you the opportunity to win a maximum reward of £20. You could be paid £10 reward for your participation on the first experiment and £10 for your participation on the second experiment.

Are there any risks involved? There are no particular risks associated with your participation other than those associated with the use of standard computer equipment.

Will my data be confidential? Any information that is obtained in connection with this study and that can be identified with you will remain confidential and will be disclosed only with your permission or as required by law. The information collected during the experiment will be kept separately from your personal identity. The information will be collected and stored on the password-protected computer where you ran the trial.

What happens if I change my mind? You may withdraw at any time and for any reason. You may access, change, or withdraw your data at any time and for any reason prior to its destruction. You may keep any benefits you receive.

What happens if something goes wrong? Should you have any concern or complaint, contact me if possible (pgg1g14@ecs.soton.ac.uk), otherwise please contact the FPSE Office (school@ecs.soton.ac.uk) or any other authoritative (RGOinfo@soton.ac.uk).

Appendix (ii) Data collection plan

Participant Information

Ethics reference number: **ERGO/**FPSE**/17155**	Version: 1	Date: 2015-08-10
Study Title: Visualisation of automation, as a tool for changing people's perception		
Investigator: Pedro Garcia Garcia, Enrico Costanza, Sarvapali Ramchurn		

Experiment 1

1. Robot A

a) How clean do you think is the room before the Robot starts to work?

1	2	3	4	5

Dirty Clean

b) How clean do you think is the room after the Robot finished with its task?

1	2	3	4	5

Dirty Clean

2. Robot B

a) How clean do you think is the room before the Robot starts to work?

1	2	3	4	5

Dirty Clean

b) How clean do you think is the room after the Robot finished with its task?

1	2	3	4	5

Dirty Clean

3. Which Robot do you think has a better performance?

Robot A ☐ Robot B ☐

Experiment 2

1. How you evaluate the performance of the systems?

System A

1	2	3	4	5
Bad				Good

System B

1	2	3	4	5
Bad				Good

System C

1	2	3	4	5
Bad				Good

2. Which system you consider that was easier for you to find the errors on the e-text?

System A ☐ System B ☐ System C ☐

3. Which system you consider that had more errors?

System A ☐ System B ☐ System C ☐

4. Which system you consider that has the best performance?

System A ☐ System B ☐ System C ☐

Appendix (iii) Consent Form template

Consent Form

Ethics reference number: **ERGO/**FPSE**/17155**	Version: 1	Date: 2015-08-10
Study Title: Visualisation of automation, as a tool for changing people's perception		
Investigator: Pedro Garcia Garcia, Enrico Costanza, Sarvapali Ramchurn		

Please initial the box(es) if you agree with the statement(s):

I have read and understood the Participant Information (version 1 dated 2015-08-10) and have had the opportunity to ask questions about the study.	☐
I agree to take part in this study.	☐
I understand my participation is voluntary and I may withdraw at any time and for any reason.	☐

Data Protection

I understand that information collected during my participation in this study is will be stored on a password protected computer and that this information will only be used in accordance with the Data Protection Act (1998).

Name of participant (print name)……

Signature of participant……..

Date……

Appendix (iv) DPA Plan template

DPA Plan

Ethics reference number: **ERGO/**FPSE**/17155**	Version: 1	Date: 2015-08-10
Study Title: Visualisation of automation, as a tool for changing people's perception		
Investigator: Pedro Garcia Garcia, Enrico Costanza, Sarvapali Ramchurn		

The following is an exhaustive and complete list of all the data that will be collected (through questionnaire and during the study):

- Participant's name
- Observations during of how the participants used the system
- Questionnaire data

Participant's name is important as the experiment will be paid for according to results. The questionnaire data will tell us how people perceive systems' performance

The data will be processed fairly, as participants will be explicitly asked about such information, and it is up to them to provide it. Any information published from this experiment will be done in aggregate or anonymous form.

The data will be processed in accordance with the rights of the participants because they will have the right to access, correct, and/or withdraw their data at any time and for any reason. Participants will be able to exercise their rights by contacting the investigator (e-mail: pgg1g14@ecs.soton.ac.uk) or the FoPSE office (e-mail: school@ecs.soton.ac.uk). Data files will be stored in a password-protected computer within the University of Southampton. Any physical data shall be destroyed, but its content will be moved to the password-protected computer.

FPSE Ethics Committee
FPSE EC Application Form Ver 6.6d

Refer to the *Instructions* and to the *Guide* documents for a glossary of the key phrases in **bold** and for an explanation of the information required in each section. The *Templates* document provides some text that may be helpful in presenting some of the required information.

Replace the highlighted text with the appropriate information.

Note that the size of the text entry boxes provided on this form does **not** indicate the expected amount of information; instead, refer to the *Instructions* and to the *Guide* documents in providing the complete information required in each section. Do **not** duplicate information from one text box to another.

Reference number: **ERGO/**FPSE**/17155**	Version: 1	Date: 2015-08-10
Name of **investigator**(s): Pedro Garcia Garcia		
Name of supervisor(s) (if student **investigator**(s)): Enrico Costanza Sarvapali Ramchurn		
Title of study: Visualisation of automation, as a tool for changing people's perception		
Expected study start date: 08/10/2015	Expected study end date: 09/25/2016	
Note that the dates requested on the "IRGA" form refer to the start and end of *data collection*. These are not the same as the start and end dates of the study for which approval is sought. Note that approval must be obtained before the study commences; retrospective approval cannot be given.		

The investigator(s) undertake to:

- Ensure the study Reference number ERGO/FPSE/17155 is prominently displayed on all advertising and study materials, and is reported on all media and in all publications;
- Conduct the study in accordance with the information provided in the application, its appendices, and any other documents submitted;
- Conduct the study in accordance with University policy governing research involving human **participants** (http://www.southampton.ac.uk/ris/policies/ethics.html);
- Conduct the study in accordance with University policy on data retention (http://www.southampton.ac.uk/library/research/researchdata/);
- Submit the study for re-review (as an amendment through ERGO) or seek FPSE EC advice if any changes, circumstances, or outcomes materially affect the study or the information given;
- Promptly advise an appropriate authority (Research Governance Office) of any adverse study outcomes, changes, or circumstances (via an adverse event notification through ERGO);
- Submit an end-of-study form as may be required by the Research Governance Office upon completion of the study.

Appendix (iii) Consent Form template

Consent Form

Ethics reference number: **ERGO/**FPSE**/17155**	Version: 1	Date: 2015-08-10
Study Title: Visualisation of automation, as a tool for changing people's perception		
Investigator: Pedro Garcia Garcia, Enrico Costanza, Sarvapali Ramchurn		

Please initial the box(es) if you agree with the statement(s):

I have read and understood the Participant Information (version 1 dated 2015-08-10) and have had the opportunity to ask questions about the study.	☐
I agree to take part in this study.	☐
I understand my participation is voluntary and I may withdraw at any time and for any reason.	☐

Data Protection

I understand that information collected during my participation in this study is will be stored on a password protected computer and that this information will only be used in accordance with the Data Protection Act (1998).

Name of participant (print name).....

Signature of participant.....

Date.....

References

Alan, A. T., Costanza, E., Ramchurn, S. D., Fischer, J., Rodden, T., and Jennings, N. R. (2016). Tariff agent: Interacting with a future smart energy system at home. *ACM Trans. Comput.-Hum. Interact.*, 23(4):25:1–25:28.

Ariely, D. (2008). *Predictably irrational*. HarperCollins New York.

Bakhshi, S., Shamma, D. A., Kennedy, L., Song, Y., de Juan, P., and Kaye, J. J. (2016). Fast, cheap, and good: Why animated gifs engage us. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 575–586, New York, NY, USA. ACM.

Balijepally, V., Nerur, S., and Mahapatra, R. (2015). Task mental model and software developers' performance: An experimental investigation. *Communications of the Association for Information Systems*, 36(1):4.

Boyce, M. W., Chen, J. Y., Selkowitz, A. R., and Lakhmani, S. G. (2015). Effects of agent transparency on operator trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, pages 179–180, New York, NY, USA. ACM.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Bretan, M., Hoffman, G., and Weinberg, G. (2015). Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human-Computer Studies*, 78:1 – 16.

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). Amazon's mechanical turk. *Perspectives on Psychological Science*, 6(1):3–5. PMID: 26162106.

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Cheema, S., VanLehn, K., Burkhardt, H., Pead, D., and Schoenfeld, A. (2016). Electronic posters to support formative assessment. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 1159–1164, New York, NY, USA. ACM.

Chuang, J., Manning, C. D., and Heer, J. (2012). “without the clutter of unimportant words”: Descriptive keyphrases for text visualization. *ACM Trans. Comput.-Hum. Interact.*, 19(3):19:1–19:29.

Detenber, B. H. and Reeves, B. (1996). A bio-informational theory of emotion: Motion and image size effects on viewers. *Journal of Communication*, 46(3):66–84.

Difallah, D. E., Catasta, M., Demartini, G., Ipeirotis, P. G., and Cudré-Mauroux, P. (2015). The dynamics of micro-task crowdsourcing: The case of amazon mturk. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 238–247, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Dittrich, W. H. and Lea, S. E. G. (1994). Visual perception of intentional motion. *Perception*, 23(3):253–268.

Dragan, A. D., Bauman, S., Forlizzi, J., and Srinivasa, S. S. (2015). Effects of robot motion on human-robot collaboration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 51–58, New York, NY, USA. ACM.

Ehrlich, K., Kirk, S. E., Patterson, J., Rasmussen, J. C., Ross, S. I., and Gruen, D. M. (2011). Taking advice from intelligent systems: The double-edged sword of explanations. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, pages 125–134, New York, NY, USA. ACM.

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., and Sandvig, C. (2015). I always assumed that i wasnt really that close to [her]: Reasoning about invisible algorithms in the news feed. In *Proceedings of the 33rd Annual SIGCHI Conference on Human Factors in Computing Systems*, pages 153–162.

Forlizzi, J. and DiSalvo, C. (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, pages 258–265, New York, NY, USA. ACM.

Fritz Heider, M. S. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2):243–259.

Gao, T., McCarthy, G., and Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, 21(12):1845–1853.

Gao, T. and Scholl, B. J. (2011). Chasing vs. stalking: interrupting the perception of animacy. *Journal of experimental psychology: Human perception and performance*, 37(3):669.

Garcia, P. G., Costanza, E., Ramchurn, S. D., and Verame, J. K. M. (2016). The potential of physical motion cues: Changing people's perception of robots' performance. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 510–518, New York, NY, USA. ACM.

Gelman, R., Durgin, F., and Kaufman, L. (1995). Distinguishing between animates and inanimates: Not by motion alone. *Causal cognition: A multidisciplinary debate*, pages 150–184.

Germine, L., Nakayama, K., Duchaine, B., Chabris, C., Chatterjee, G., and Wilmer, J. (2012). Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19(5):847–857.

Healey, C. G. and Enns, J. T. (1999). Large datasets at a glance: combining textures and colors in scientific visualization. *IEEE Transactions on Visualization and Computer Graphics*, 5(2):145–167.

Herlocker, J. L., Konstan, J. A., and Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 241–250, New York, NY, USA. ACM.

Hoffman, G. and Vanunu, K. (2013). Effects of robotic companionship on music enjoyment and agent perception. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, HRI '13, pages 317–324, Piscataway, NJ, USA. IEEE Press.

Horvitz, E. and Krumm, J. (2012). Some help on the way: Opportunistic routing under uncertainty. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 371–380, New York, NY, USA. ACM.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211.

Jung, J., Bae, S.-H., and Kim, M.-S. (2013a). Three case studies of with moving products. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 509–518, New York, NY, USA. ACM.

Jung, J., Bae, S.-H., Lee, J. H., and Kim, M.-S. (2013b). Make it move: A movement design method of simple standing products based on systematic mapping of torso movements & product messages. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 1279–1288, New York, NY, USA. ACM.

Kato, J., Sakamoto, D., Igarashi, T., and Goto, M. (2014). Sharedo: To-do list interface for human-agent task sharing. In *Proceedings of the Second International Conference on Human-agent Interaction*, HAI '14, pages 345–351, New York, NY, USA. ACM.

Kazai, G., Kamps, J., and Milic-Frayling, N. (2013). An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178.

Kieras, D. E. and Bovair, S. (1984). The role of a mental model in learning to operate a device*. *Cognitive Science*, 8(3):255–273.

Kim, T. and Hinds, P. (2006). Who should i blame? effects of autonomy and transparency on attributions in human-robot interaction. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*, pages 80–85.

Lim, B. Y. and Dey, A. K. (2011a). Design of an intelligible mobile context-aware application. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 157–166, New York, NY, USA. ACM.

Lim, B. Y. and Dey, A. K. (2011b). Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 415–424, New York, NY, USA. ACM.

Lim, B. Y., Dey, A. K., and Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 2119–2128, New York, NY, USA. ACM.

Lyons, J. (2013). Being transparent about transparency: A model for human-robot interaction.

Mason, W. and Watts, D. J. (2010). Financial incentives and the "performance of crowds". *SIGKDD Explor. Newsl.*, 11(2):100–108.

Mateas, M. (1999). *An Oz-Centric Review of Interactive Drama and Believable Agents*, pages 297–328. Springer Berlin Heidelberg, Berlin, Heidelberg.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Michotte, A. (1963). The perception of causality.

Mortensen, D. H., Hepworth, S., Berg, K., and Petersen, M. G. (2012). "it's in love with you": Communicating status and preference with simple product movements. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, pages 61–70, New York, NY, USA. ACM.

Norman, D. A. (2013). *The design of everyday things: Revised and expanded edition*. Basic books.

Nowacka, D., Hammerla, N. Y., Elsden, C., Plötz, T., and Kirk, D. (2015). Diri - the actuated helium balloon: A study of autonomous behaviour in interfaces. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 349–360, New York, NY, USA. ACM.

O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., and Höllerer, T. (2008). Peerchooser: Visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1085–1088, New York, NY, USA. ACM.

Pantelis, P. C. and Feldman, J. (2012). Exploring the mental space of autonomous intentional agents. *Attention, Perception, & Psychophysics*, 74(1):239–249.

Park, D. and Lee, J.-H. (2010). *Investigating the Affective Quality of Motion in User Interfaces to Improve User Experience*, pages 67–78. Springer Berlin Heidelberg, Berlin, Heidelberg.

Popović, J., Seitz, S. M., and Erdmann, M. (2003). Motion sketching for control of rigid-body simulations. *ACM Trans. Graph.*, 22(4):1034–1054.

Reeves, B. and Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press Cambridge, UK.

Richard, G. (1998). Brainy mind. *British Medical Journal*, 137:1693–1695.

Rousseau, D. and Hayes-Roth, B. (1997). Interacting with personality-rich characters. *Report No. KSL 97*, 6.

Schacter, D. L., Eich, J. E., and Tulving, E. (1978). Richard semon's theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 17(6):721 – 743.

Schlottmann, A. and Surian, L. (1999). Do 9-month-olds perceive causation-at-a-distance? *Perception*, 28(9):1105–1113.

Schulz, D., Burgard, W., Fox, D., Thrun, S., and Cremers, A. (2000). Web interfaces for mobile robots in public places. *Robotics Automation Magazine, IEEE*, 7(1):48–56.

Schwarz, N. and Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of personality and social psychology*, 45(3):513.

Takayama, L., Dooley, D., and Ju, W. (2011). Expressing thought: Improving robot readability with animation principles. In *Proceedings of the 6th International Conference on Human-robot Interaction*, HRI '11, pages 69–76, New York, NY, USA. ACM.

Tremoulet, P. D. and Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, 29(8):943–951.

Tremoulet, P. D. and Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Perception & psychophysics*, 68(6):1047–1058.

Tullio, J., Dey, A. K., Chalecki, J., and Fogarty, J. (2007). How it works: A field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 31–40, New York, NY, USA. ACM.

Tversky, A. and Kahneman, D. (1985). The framing of decisions and the psychology of choice. In *Environmental Impact Assessment, Technology Assessment, and Risk Analysis*, pages 107–129. Springer.

Verame, J. K. M., Costanza, E., and Ramchurn, S. D. (2016). The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4908–4920, New York, NY, USA. ACM.

Verbert, K., Parra, D., Brusilovsky, P., and Duval, E. (2013). Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 351–362, New York, NY, USA. ACM.

Vermeulen, J. (2010a). Improving intelligibility and control in ubicomp. In *Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing - Adjunct*, UbiComp '10 Adjunct, pages 485–488, New York, NY, USA. ACM.

Vermeulen, J. (2010b). Improving intelligibility and control in ubicomp. In *Proceedings of the 12th ACM International Conference Adjunct Papers on Ubiquitous Computing - Adjunct*, UbiComp '10 Adjunct, pages 485–488, New York, NY, USA. ACM.

Vermeulen, J., Luyten, K., and Coninx, K. (2013). Intelligibility required: How to make us look smart again.

Vines, B. W., Krumhansl, C. L., Wanderley, M. M., and Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition*, 101(1):80 – 113.

Wainer, J., Feil-Seifer, D. J., Shell, D. A., and Mataric, M. J. (2007). Embodiment and human-robot interaction: A task-based perspective. In *Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on*, pages 872–877.

Ware, C. (2012). *Information visualization: perception for design*. Elsevier.

Woods, S., Walters, M., Koay, K. L., and Dautenhahn, K. (2006). Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In *9th IEEE International Workshop on Advanced Motion Control, 2006.*, pages 750–755.

Yang, R. and Newman, M. W. (2012). Living with an intelligent thermostat: Advanced control for heating and cooling systems. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pages 1102–1107, New York, NY, USA. ACM.

Yang, R. and Newman, M. W. (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, pages 93–102, New York, NY, USA. ACM.

Yatani, K., Novati, M., Trusty, A., and Truong, K. N. (2011). Review spotlight: A user interface for summarizing user-generated reviews using adjective-noun word pairs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 1541–1550, New York, NY, USA. ACM.

Youngblood, G. M. and Cook, D. J. (2007). Data mining for hierarchical model creation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(4):561–572.