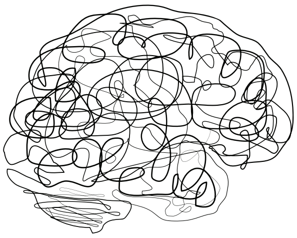

PROCEEDINGS OF THE 17TH SWECOG CONFERENCE

Örebro 2022

16 - 17 June

Editors

Hadi Banaee
Erik Billing

SweCog
SWEDISH COGNITIVE SCIENCE SOCIETY

Editors:
Hadi Banaee, Örebro University, hadi.banaee@oru.se
Erik Billing, University of Skövde, erik.billing@his.se

www.swecog.se

Skövde University Studies in Informatics 2022:1
ISBN 978-91-983667-9-2
ISSN 1653-2325

PUBLISHED BY THE UNIVERSITY OF SKÖVDE

Preface

Welcome to SweCog 2022 in Örebro!

This booklet contains the abstracts and short papers for all oral and poster presentations at the 2022 SweCog conference. This year, SweCog 2022 has been held jointly with the international workshop on Artificial Intelligence and Cognition (AIC).

Following the SweCog tradition, with the aim to support networking among researchers in Sweden, contributions cover a wide spectrum of cognitive science research. However, as a result of that the conference has been arranged jointly with AIC, a considerable portion of the contributions this year addresses the problems in both fields of cognitive science and artificial intelligence. More specifically, there are a number of contributions that discuss the trending topic of ”Human-Centered AI” which was also the focusing topic of the AIC workshop. Within the works addressing AI and cognition, the key concept of ”learning” is at the centre of focus. In these studies, the contributions consider studying the human learning processes from the cognitive point of view, rather than building artificial models of learning (as called ”machine learning” in AI).

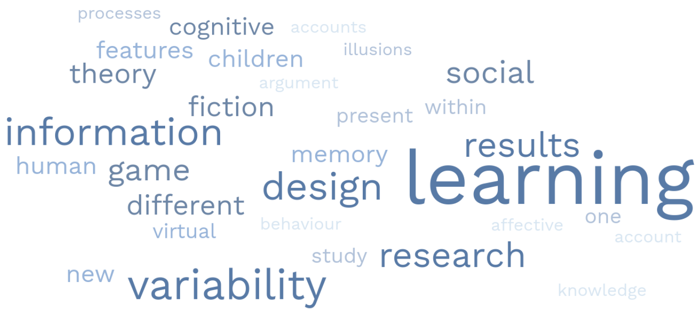

Figure 1: The word cloud of the terms used in the abstracts of the accepted papers.

Hadi Banaee and Erik Billing

The reviewers were:

Beatrice Alenljung, Erik Billing, Andreas Falck, Pierre Gander, Linus Holm, Erik Lagerstedt, Maurice Lamb, Julia Rosén, Jana Rambusch, and Sofia Thunberg.

Conference Programme

Parts of SweCog 2022 runs jointly with the *8th International Workshop on Artificial Intelligence and Cognition (AIC)*. Please refer to the AIC program for a full list of speakers.

Thursday June 16th

08:30 — 09:00	*Registration*
09:00 — 10:00	Invited speaker — **Ute Schmid** (p. 5)
10:30 — 12:00	Oral presentations by Andreas Kalckert (p. 10) and others, *shared with AIC.*
13:15 — 14:10	Oral presentations by Joel Parthemore (p. 28) and others, *shared with AIC.*
14:10 — 16:00	**Poster session** hosting short presentations by Oscar Bjurling (p. 7), Emma Mainza Chilufya (p. 16), Andreas Falck (p. 7), Philip Gustafsson (p. 9), Maybí Morell Ruiz (p. 10), Anders Persson (p. 11), Anna Persson (p. 12), and Alexander Tagesson (p. 12).
16:00 — 17:00	Invited speaker - **Kees van Deemter** (p. 5)

Friday June 17th

09:00 — 10:15	Oral presentations by Pierre Gander (p. 8), Leonard Ngaosuvan (p. 11), and Linus Holmberg (p. 9).
10:45 — 11:50	Oral session shared with AIC. Please refer to the AIC program for speaker details.
13:15 — 14:45	Oral presentations by Andreas Falck (p. 20), Amandus Krantz (p. 25), and Raphaël Fargier (p. 8).
15:20 — 15:30	*Conference closing*

Keynote presentations

Hybrid, Explanatory, Interactive Machine Learning– Towards Trustworthy Human-AI Partnerships

Prof. Ute Schmid

For many practical applications of machine learning, it is appropriate or even necessary to make use of human expertise to compensate a too small amount or low quality of data. Taking into account knowledge which is available in explicit form reduces the amount of data needed for learning. Furthermore, even if domain experts cannot formulate knowledge explicitly, they typically can recognize and correct erroneous decisions or actions. This type of implicit knowledge can be injected into the learning process to guide model adaptation. These insights have led to the so-called third wave of AI with a focus on explainability (XAI). In the talk, I will introduce research on explanatory and interactive machine learning. I will present inductive programming as a powerful approach to learning interpretable models in relational domains. Arguing the need for specific explanations for different stakeholders and goals, I will introduce different types of explanations based on theories and findings from cognitive science. Furthermore, I will show how intelligent tutor systems and XAI can be combined to support constructive learning. Algorithmic realisations of explanation generation will be complemented with results from psychological experiments investigating the effect on joint human-AI task performance and trust. Finally, current research projects are introduced to illustrate applications of the presented work in medical diagnostics, quality control in industrial production, file management, and accountability.

Explanation and Rationality in Models of Language

Prof. Kees van Deemter

When theories of human behaviour aim to offer explanations, they often use rationality as their linchpin: to the extent that a theory helps us to see behaviour as optimising some form of rationality/utility, we feel that our theory explains this behaviour. This approach is not uncontroversial, however. For example, four decades of research in Behavioural Economics have shown that people behave in ways that are not easily explained by rationality alone.

Rationality has long had its adherents in the explanation of language use as well, for example via the Gricean Maxims. Recently, a Bayesian approach known as Rational Speech Act (RSA) theory has made inroads into the computational modelling of language use. In a nutshell, the idea is to build tightly coupled models of language comprehension and production in which speakers and hearers assume each other to behave rationally.

In this talk I will sketch a series of experiments focussing on the way in which speakers refer to objects. These experiments paint a less "rational" picture of human language use, and they offer confirmation of a model, known as Probabilistic Referential Overspecification (PRO), that balances rationality with other considerations. I hope to engage in a discussion of the dilemma of having to choose between two these very different models, one of which is elegant and explanatory yet empirically inadequate, while the other is messy yet empirically very adequate.

Abstracts

Human Interaction with Autonomous Drone Swarms: Design and control challenges

Oscar Bjurling

To date, Human-Swarm Interaction (HSI) research has largely focused on different problem areas in isolation, missing potential interaction effects between drone swarm architecture designs, control methods, and user interfaces that impacts system (and interaction) complexity. This highlights the pressing need for a holistic research approach. There is also a need for work-driven (complementing technology-driven) research to ensure the usability of swarm systems. Therefore, the current research project explores how swarm systems can assist in actual work environments (like forest firefighting or maritime search and rescue (SAR) operations), what capabilities they require, what challenges they pose to their operators, and how to design useable and efficient human-swarm interfaces for these work contexts. The conceptual work carried out in the current research project suggests that, in a real work context, a (de)centralized hybrid-control approach is required to strike a balance between swarm autonomy and resilient operation on the one hand, and operator control, situational awareness, and mental workload on the other. The system and user interface design must allow for the traversal between and within system strata, ranging from swarms and subswarms to individual drones and their sensors or equipment. For instance, this is important to a SAR swarm operator who must delegate tasks and supervise individual and groups of drones during a mission. Other current project results suggests that the swarm metaphor is, in fact, antithetical to the mission and user requirements presented above, and that choir or orchestra are perhaps better metaphors for generating useful designs.

The leader learns it all? Using the “Kaptein Morf” tablet game to examine how different roles in joint problem solving affect learning

Andreas Falck and Janne von Koss Torkildsen

Kaptein Morf is a morphology-based vocabulary learning game for children aged 7-9, as well as a research tool (Torkildsen et al. 2021). Here, we describe the conceptual design of a multiplayer extension to the game, in which children can solve tasks while jointly attending each other's solutions. The scientific aim of the multi-player game is to address how different roles within the joint attentional exchanges affect learning outcomes. For some tasks within the game session, children will be assigned to be the "leader", i.e. initiating a response and selecting the final solution, and for other tasks to be the "follower", i.e. having only an advisory role in the problem-solving. Learning in these two conditions will be compared to a baseline of solo play. The game allows tracking of learning on the level of single task items, enabling within-subject manipulation of the "leading" and "following" roles within the same game session. This makes the game engaging for the children, while still maintaining precise experimental control of the children's turn-taking. The multiplayer extension is currently in development, and the present poster demonstrates how joint attention will be implemented in the game, with focus on the roles as "leader" and "follower".

The influence of contextual variability on learning novel words: Does the type of variability matter?

Raphaël Fargier, Andreas Falck, Tine Hovland,
Hakan Bayar, and Janne von Koss Torkildsen

Adults predominantly learn new vocabulary from reading, and contextual variability benefits such learning. Contextual variability often refers to the number of unique documents a new word appears in, or to the number of different topics covered by the texts. Additionally, visual variability has been found to benefit learning of object words in children. In particular, variability in irrelevant object features (e.g. presenting chairs in different colors and materials) help children determine the core features of the object (e.g. that the core feature of a chair is its shape, not its color or material). In the present study, we examine what features of variability facilitate learning of novel object words from narrative contexts. We manipulated variability in non-definitional object features (e.g. color, size) and variability in situational contexts in which new objects are experienced (e.g. characters, location). In web-based experiments, participants encountered novel words in blocks of three short fictional narratives, and then provided a written definition of that word. Pilot data showed that lexical recognition performance was at ceiling at the immediate test, still high a week later (follow-up), and better in the condition with variability in non-definitional object features. Definition scores indicated better learning of the core semantic features in the condition with the highest degree of variability, i.e. variability in both non-definitional and situational features. Results suggest that situational variability may hinder lexical retention but may support better identification of core semantic features.

What kind of memory is memory of fiction?

Pierre Gander

Much of information people encounter in everyday life is not factual, such as from movies, novels and computer games. In recent years, there has been an increase in research on fiction, but memory of fiction and effects of fiction has been treated as isolated phenomena. There is a need for a theoretical account of how memory of fictional information is related to other types of memory and which mechanisms allow people to separate fact and fiction in memory. In this theoretical work, we propose an extension of Rubin's dimensional memory model to account for memories of fictional information of events, places, characters, and objects. Further, we offer a set of proposed mechanisms involving various degrees of complexity and levels of conscious processing, that mostly keep fact and fiction separated, but also allow learning and misinformation from fiction: content-based reasoning, source monitoring, and an associative link from the memory to the concept of fiction. In this way, we characterize the processing of fiction as a fundamental cognitive process that is innate, culturally universal, spontaneous, and independent of medium and modality and whether the information is mediated or directly experienced.

Vocal Characteristics predict Accuracy in Eyewitness Testimony

Philip Gustafsson

In two studies, we examined if correct and incorrect testimony statements were produced with vocally distinct characteristics. Participants watched a staged crime film and were interviewed as eyewitnesses. Witness responses were recorded and analyzed along 16 vocal dimensions. A mega-analysis of the two datasets showed four distinct vocal characteristics of accuracy; correct responses were uttered with a higher pitch, a ”fuller voice”, higher speech rate and shorter pauses. Taken together, this study advances previous knowledge by showing that accuracy is not only indicated by what we say, but also by how we say it.

Sexual Economics in Swedish Dating: Pity Poor Men

Linus Holmberg

Sexual exchange theory (SET) is a controversial theory describing heterosexual partner selection in terms of economic market factors. This paper explores SET empirically in Sweden, one of the most financially equal nations in the world. Experiment 1, a vignette study with four dating profiles, tested whether access to resources increases male attractiveness. Experiment 2, a vignette study measured how justifiable men's disappointment was, depending on financial courtship investments in a failed courtship attempt. The results of Experiment 1 indicated that, even in Sweden, men with limited resources are considered less attractive. Male financial resources are not seen as a bonus, but rather a prerequisite. In Experiment 2, participants felt that it was not justifiable to be disappointed for men who were ‘cheap’ in courtship. These results indicate that SET is a useful theory, even in a relatively gender-equal society.

From rubber hands to virtual hands – A critical examination of the processes underlying bodily illusions

Andreas Kalckert

Bodily illusions such as the rubber hand illusion are well-known paradigms within experimental psychology and cognitive neuroscience. These paradigms have gained in popularity, with new variations of the illusions introduced almost every year. These new variants may include different sensory information (e.g., movements instead of tactile input) or other manipulations of the body (e.g., shape or look of the hand). Likewise, these illusions have been deployed in virtual reality which allows further manipulations not permitted in real settings. Most researchers draw equivalencies between these different variations, concluding for example that the illusion in virtual reality works in a similar manner to the real setting. In this talk I like to highlight two general problems with these interpretations: first, the assumptions of the underlying perceptual and cognitive processes generating the illusion experience and second, the way these illusions are quantified and results are interpreted. I like to point to certain caveats in bodily illusion paradigms this way. These issues may be important to consider in future applications of bodily illusions.

What do our eyes say about our estimation strategies?

Maybí Morell Ruiz, Magnus Haake, and Agneta Gulz

Numerical estimation, measured with the Number Line Estimation Task (NLET), has been related to mathematical competence [Schneider et al.] and numerical knowledge development [Siegler, 2022]. In this, eye-tracking has shown promising results in developmental studies of number sense [Schneider et al., 2008] and knowledge of numerical magnitude [Heine et al., 2010] in children. Combining embedded eye-tracking technologies in laptops and tablets, preschoolers’ development of numerical estimation can be evaluated and integrated in early math educational software by means of pedagogically adaptive algorithms [Gulz et al., 2022]. In this pilot study with 10 PhD-students, performance (AEE, Absolute Estimated Error) and estimation strategies (eye fixation patterns) were evaluated using a laptop setup with eye-tracking and an on-screen implementation of NLET in a bounded and unbounded condition. Results on performance show that the unbounded condition (M=8.9; SD=5.92) has a lager AEE than the bounded condition (M=4.6; SD=2.43), with a significant medium effect size difference between conditions (t(26)=3.40, p¡.002, Cohen’s d=0.65). Results for the estimation strategies replicate previous findings [Reinert et al., 2015], with eye-fixation patterns in the bounded condition describing a W shape and the unbounded a systematic downward trend. A next step is to embed this NLET eye-tracking methodology in a play-and-learn game for preschoolers.

Cognitive bias in social services CPS case argumentation

Leonard Ngaosuvan

Child protective service (CPS) cases concern taking children into protective custody. Generally, social services investigate and present arguments for protective custody in court. The present study investigated a new type of cognitive bias in social services CPS case argumentation. The bias was first detected in an actual CPS investigation. The present study investigated the external validity of the bias. Participants (N = 133) completed an online within-subjects experiment where they rated the plausibility of two illogical arguments' (Simple vs. Complex), and six distractor items. The simple argument was as follows: "A is taller than B, hence C is taller than B". The complex argument was an abbreviated version of the actual CPS case where the parents appeared to provide inadequate attachment with the child. Broken down, the complex argument had the same isomorphic structure as the simple argument. The results showed that complex argument was considered implausible by 53%, and the simple 79%. The same pattern was found among participants with relevant academic training (N = 42); social worker, lawyer, psychologist, and students of said topics), 52% and 83% respectively. The results are discussed in terms of a new cognitive bias, and cognitive overload.

What is Reason in the age of Artificial Intelligence and predictive processing?

Anders Persson

From philosophical and scientific accounts dating back as long as the Ancients Greeks, upon until contemporary days, there are on the face of it similar dyadic accounts for what Human Reasoning is, and is not. For Plato and Aristotle, Reason was an intellectual activity aimed for truth and knowledge, limited to Intellectuals and contrasted with Workers aiming to satisfy their Desire. Similar distinctions can be found with Kant and Hume, being in the Sensible or Intelligible worlds, aiming to satisfy Thinking, or Passions. Entering the 20th-centurary you have the distinction between Intuition and Reason, or some kind of Critical Thinking, and more recent accounts such as Implicit and Explicit knowledge and processes, or the infamous System 1 and System 2. With Artificial Intelligence on the agenda it begs the question, more than ever, what makes up a cognitive, intelligent, reflecting, thinking and reasoning agent, and it is not all that clear what previous accounts make out of a "Reasoning process". Accepting a distinction between Intuition and Reason, there is an interesting question what predictive processing as an account of the brain adds to—with a generative working model simulated onto the world, and corrected by errors, it may seem to mostly be about Intuition. But coupled with, so called, "Offline" mental simulations, it might be an interesting account for Reason, that also might account for common critique of Human

Reasoning abilities. In the poster presentation I will try to account for these theoretical considerations, as part of an ongoing development with my thesis.

The role of prior experience in understanding speech: computational and experimental approaches to vowel perception

Anna Persson

One of the central challenges for speech perception is that talkers differ in their pronunciations. This results in between-talker differences in the mapping from the acoustic signal onto linguistic categories and meanings [Liberman et al., 1967]. Yet, listeners are remarkably adept at overcoming the initial difficulty in understanding new talkers [e.g. Clarke and Garrett, 2004, Xie et al., 2017]. Despite substantial progress, the mechanisms that underlie these adaptive abilities remain unknown. I will present the initial steps of my research on this question for the perception of English vowels. I develop computational models [ideal observers, see e.g. Feldman et al., 2009, Kleinschmidt and Jaeger, 2015] based on phonetic databases, and test their predictions in web-based perception experiments to investigate whether listeners learn and store talker- and group-specific phonetic representations [e.g. Pierrehumbert, 2001], and how the answer to this question might depend on pre-linguistic normalization procedures [like those commonly used in phonetic research, e.g. Lobanov, 1971, Nearey, 1978, McMurray and Jongman, 2011].

Do objective judges become emotional?

Alexander Tagesson

Affective processes are an integral part of much of juridical decision-making. Several researchers claim that affective processes, such as empathy and compassion, are parochial and biased, creating inconsistent decision-making [Bloom, 2016, Slovic, 2007, Cameron et al., 2019]. Consistency becomes especially important in juridical contexts, where inconsistent decisions can undermine the rule of law [e.g., Pärnamets et al., 2020].

With this background in mind, we tested how affective information, e.g. characterizing someone in a positive or negative way, and affective processes, e.g. how much compassion or empathy was felt with someone, affected Swedish district court judges' decision-making during remand proceedings. The judges were asked to make several decisions connected to different applications for remand orders. Cases were presented as short vignettes, paired with a picture of the defendant and were designed to resemble real remand proceedings. Specific cases were matched on judicially relevant information, but, importantly, were mismatched on affective information. This design allowed us to examine how affective information affect juridical decision-making. We used self-reported empathy towards defendants and victims to predict judicially relevant decision outcomes.

Bibliography

Paul Bloom. *Against empathy: The case for rational compassion*. Ecco press, 1st editio edition, 2016. ISBN 0062339338. doi: 10.1097/TA.0000000000002071. URL https://books.google.de/books?hl=de&lr=&id=_eslDAAAQBAJ&oi=fnd&pg=PT2&dq=Against+Empathy:+The+Case+for+Rational+Compassion&ots=gZI9DTFEd8&sig=VPUBGwwqegs9DS7UzpHJojT1dKs#v=onepage&q=AgainstEmpathy%3ATheCaseforRationalCompassion&f=false.

C. Daryl Cameron, Cendri A. Hutcherson, Amanda M. Ferguson, Julian A. Scheffer, Eliana Hadjiandreou, and Michael Inzlicht. Empathy is hard work: People choose to avoid empathy because of its cognitive costs. *Journal of Experimental Psychology: General*, 148(6):962–976, 6 2019. ISSN 00963445. doi: 10.1037/XGE0000595.

Constance M. Clarke and Merrill F. Garrett. Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, 116(6):3647–3658, December 2004. ISSN 0001-4966. doi: 10.1121/1.1815131.

Naomi H. Feldman, Thomas L. Griffiths, and James L. Morgan. The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4):752–782, October 2009. ISSN 1939-1471, 0033-295X. doi: 10.1037/a0017196.

Agneta Gulz, Ludde Londos, and Magnus Haake. Preschoolers' understanding of a teachable agent-based game in early mathematics as reflected in their gaze behaviors – an experimental study. *International Journal of Artificial Intelligence in Education*, 30:30–73, 2022.

Angela Heine, Verena Thaler, Sascha Tamm, Stefan Hawelka, Michael Schneider, Joke Torbeyns, Bert De Smedt, Lieven Verschaffel, Elsbeth Stern, and Arthur M Jacobs. What the eyes already 'know': using eye movement measurement to tap into children's implicit numerical magnitude representations. *Infant and Child Development: An International Journal of Research and Practice*, 19(2):175–186, 2010.

Dave F. Kleinschmidt and T. Florian Jaeger. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122 (2):148–203, April 2015. ISSN 1939-1471, 0033-295X. doi: 10.1037/a0038695.

A. M. Liberman, F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. Perception of the speech code. *Psychological Review*, 74(6):431–461, 1967. ISSN 1939-1471, 0033-295X. doi: 10.1037/h0020279.

B. M. Lobanov. Classification of russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, 49(2B):606–608, February 1971. ISSN 0001-4966. doi: 10.1121/1.1912396.

Bob McMurray and Allard Jongman. What information is necessary for speech categorization? harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, 118(2):219–246, 2011. ISSN 1939-1471, 0033-295X. doi: 10.1037/a0022325.

T.M. Nearey. *Phonetic Feature Systems for Vowels.* PhD thesis, 1978.

Philip Pärnamets, Alexander Tagesson, and Annika Wallin. Inconsistencies in repeated refugee status decisions. *Journal of Behavioral Decision Making*, 33(5):569–578, 12 2020. ISSN 1099-0771. doi: 10.1002/BDM.2176. URL https://onlinelibrary.wiley.com/doi/full/10.1002/bdm.2176https://onlinelibrary.wiley.com/doi/abs/10.1002/bdm.2176https://onlinelibrary.wiley.com/doi/10.1002/bdm.2176.

Janet B. Pierrehumbert. Exemplar dynamics: Word frequency, lenition and contrast., 2001.

Regina M Reinert, Stefan Huber, Hans-Christoph Nuerk, and Korbinian Moeller. Strategies in unbounded number line estimation? evidence from eye-tracking. *Cognitive Processing*, 16:359–363, 2015.

Michael Schneider, Simon Merz, Johannes Stricker, Bert De Smedt, Joke Torbeyns, Lieven Verschaffel, and Koen Luwel. Associations of number line estimation with mathematical competence: A meta-analysis. *Child Development*, (5):1467–1484.

Michael Schneider, Angela Heine, Verena Thaler, Joke Torbeyns, Bert De Smedt, Lieven Verschaffel, Arthur M Jacobs, and Elsbeth Stern. A validation of eye movements as a measure of elementary school children's developing number sense. *Cognitive Development*, 23(3):409–422, 2008.

Robert S. Siegler. Development of numerical knowledge. pages 361–382, 2022.

Paul Slovic. If I Look at the Mass I Will Never Act. In *Judgment and Decision Making*, volume 2(A), page 79–95. Springer Science and Business Media B.V., 2007. doi: 10.1007/978-90-481-8647-1{_}3. URL https://link.springer.com/chapter/10.1007/978-90-481-8647-1_3.

Xin Xie, Rachel M. Theodore, and Emily B. Myers. More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, 43(1):206–217, January 2017. ISSN 1939-1277, 0096-1523. doi: 10.1037/xhp0000285.

The Design of Intelligent Virtual Agents Using User-Centered Design Methods

Emma Mainza Chilufya[1]

[1] *Department of Computer and Information Science, Linköping University, Sweden*
emma.chilufya@liu.se

1 Introduction

This paper outlines my PhD thesis project about the design of intelligent virtual agents (IVA). Ferbs [6] defines an IVA as "a physical or virtual entity that can act, perceive its environment (in a partial way) and communicate with others, is autonomous and has skills to achieve its goals and tendencies". IVAs have potential applications in many *shared spaces*, such as first-line customer support, guiding in museums, receptions, etc. The design of IVAs is multidisciplinary and focuses on different user-centred aspects such as presence, emotion, appearance, behaviour and dialogue. Yet, design choices regarding these aspects are often based on the "introspective examination of personal preferences"Isbister et al.[7] rather than any accurate reflection of the design goals or the qualities valued by the users.

2 Survey: State of the Art

In 2018 Norouzi et al [10] presented a systematic review of user studies published at the IVA conferences from 2001-2015. They showed that from 2001 to 2010 the number of user studies increased tenfold, 247 out of 579 papers described user studies. The reported studies provide important insights into various aspects of users' perceptual, behavioural and cognitive responses to virtual agents, as gathered through experiments. Though these studies give important insights and general knowledge on how to model various aspects of agents, they do not easily transform into guidelines on how to create specific agents [4]. Other studies have shown that different user groups, for example children and elderly people have different preferences when it come to agent appearance [14]. We also know that culture and the application domain are important factors for interaction style [15, 8].

3 Thesis Problem/Question

There currently seem to be no standard methodologies in Virtual Agents research that focus on the involvement of users during the design phase. The design of IVAs tends to focus on the specific aspects (rather than the IVA in it's entirety) and cannot be easily transformed into guidelines. Users are usually involved during the evaluation phase.

The aim of this thesis is to define recommendations for how to use User-Centered Design (UCD) methods in the design of IVAs for shared spaces, not only for evaluation, but also in concept generation and prototyping stages of the design process. Shared spaces in this case can be seen as a shift of focus toward supporting the context within which interaction with the IVA takes place. A space that spans the dimensions of a physical and synthetic environment [1]. The overarching research questions are:

- How might we do UCD of IVAs for shared spaces?
  - What current processes and methods are used to design virtual agents? To what extent and how are users involved in the design process of virtual agents?
  - What are the suitable ways of doing UCD of IVAs? What are the benefits and drawbacks of the methods?

4 Method

The first part was to explore the current methods used to design intelligent virtual agents. This involved a systematic review of the last 5 years of papers from the Intelligent Virtual Agent (IVA) Conference. The IVA conference is the largest in the IVA field and primarily focuses on the design and development of intelligent virtual agents in all aspects. This ranges from neuroscience to machine learning, dialogue, motion, emotion.

The second part involved the use of a case study to explore the thesis problem. The first case study looked at the design of an Intelligent Virtual Receptionist of a university department. The case was divided into two phases (conceptual and prototyping). The conceptual phase involved two workshops: virtual bodystorming with members of staff from the department, and remote desktop walkthrough with university students. The design process began from the conceptual phase as it at this phase where it is decided on *what* should be designed and *why*. Early user involvement is beneficial for usability and user experience. It brings attention to practical functionality and how the system fits the context of use [9].

Bodystorming is a form of brainstorming using participants' bodily presence on the context of use to gain insight into the user experience [11, 13]. It takes advantage of embodied cognition and interaction as embodied design methods enable the use of all of a person's senses in an emergent design space [17]. The bodystorming workshop was carried out in virtual reality with a 3D model of the office building as the environment.

Desktop walkthrough allows for a quick simulation of a service experience using simple small figurines such as LEGO pieces to represent people or other elements of service [2, 13]. To emulate that, the desktop walkthrough was carried out in Miro and a combination of LEGO and other figurine representations, to achieve a look and feel that would be similar to an ordinary face-to-face desktop walkthrough.

The prototyping phase is of a multi-platform virtual receptionist which is based on the results of the conceptual phase. This will be followed by the evaluation of the prototype with the users at the university department.

5 Results

In the case of the systematic review of the last 5 years of papers from the IVA Conference, 14% of the publications indicate some form of user involvement during design. 8%(23 papers) explicitly mentioned user involvement and have details on the users and how they were involved. 9 of the 23 papers include one-time user involvement (at the initial stage of design) and 10 papers indicated iterative involvement. Details of the evaluation process can be found in the paper Chilufya and Silvervarg [5].

The ideas generated during the bodystorming and desktop walkthrough were structured into a Morphological Chart [12, 16]. The Morphological Chart structure is based on Burk's Pentad of human actions and motives [3]. With that, three design concepts (one main and two alternative) were created [4]. The concepts were created from the morphological chart using the following criteria:

- feasibility—is it feasible to design and implement?
- desirability—is it desirable from a user's point of view?
- novelty—is it interesting and original?

The main concept is a cross-platform virtual receptionist that provides information to all human agents through different media in a user journey across the user's mobile device, a large screen, and a physical robot. The concept is based on ideas that surfaced in both workshops. The second concept is a mystical (ghost-like) virtual receptionist. The receptionist provides details on the availability of members of staff to students (human agent) and allows students to book time slots on the members of staff's schedules. The receptionist is available in specified locations and can be accessed using a student card.

The third concept is a schedule custodian virtual receptionist that assists members of staff manager their schedule and room bookings. The receptionists helps enhance conversation in the coffee area as well. Detailed results of the conceptual phase can be found in the paper Chilufya and Arvola [4].

6 Conclusion

The systematic review shows that the IVA community mostly develop interactive agents without articulating the design methods employed. With very few studies that mention design details [5]. One hypothesis is that the design of some aspects does not need the involvement of users. In some cases, users might only be required during the evaluation phase. The design details could also be published elsewhere, or are not published at all [5].

The case study presents a combination of UCD methods that are novel in the area of IVA design. The work combines embodied but remote methods with morphological chart [4]. A working hypothesis is that bodystorming yields more aesthetically focused ideas about embodied interaction while the desktop walkthrough gives a more instrumental usability focus [4].

The concept of a cross-platform IVA is interesting for further research and prototypes are currently being created. This is followed up by the second case study which will look at the design of a IVA to help young students find interest in reading books..

References

[1] S. Benford, C. Greenhalgh, G. Reynard, C. Brown, and B. Koleva. Understanding and constructing shared spaces with mixed-reality boundaries. *ACM Transactions on computer-human interaction (TOCHI)*, 5(3):185–223, 1998.

[2] J. Blomkvist, A. Fjuk, and V. Sayapina. Low threshold service design: desktop walkthrough. In *Service Design Geographies. Proceedings of the ServDes. 2016 Conference*, number 125, pages 154–166. Linköping University Electronic Press, 2016.

[3] K. Burke. *A grammar of motives*, volume 177. Univ of California Press, 1969.

[4] E. M. Chilufya and M. Arvola. Conceptual designing of a virtual receptionist: Remote desktop walkthrough and bodystorming in VR. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, pages 112–120, 2021.

[5] E. M. Chilufya and A. Silvervarg. The black box of virtual agent design: A literature review of user involvement at the iva conference. In *3rd African Human-Computer Interaction Conference: Inclusiveness and Empowerment*, pages 146–150, 2021.

[6] J. Ferber and G. Weiss. *Multi-agent systems: an introduction to distributed artificial intelligence*, volume 1. Addison-Wesley Reading, 1999.

[7] K. Isbister and P. Doyle. Design and evaluation of embodied conversational agents: A proposed taxonomy. In *The first international joint conference on autonomous agents & multi-agent systems.* Citeseer, 2002.

[8] T. Koda, T. Hirano, and T. Ishioh. Development and perception evaluation of culture-specific gaze behaviors of virtual agents. In *International Conference on Intelligent Virtual Agents*, pages 213–222. Springer, 2017.

[9] S. Kujala. User involvement: a review of the benefits and challenges. *Behaviour & information technology*, 22(1):1–16, 2003.

[10] N. Norouzi, K. Kim, J. Hochreiter, M. Lee, S. Daher, G. Bruder, and G. Welch. A systematic survey of 15 years of user studies published in the intelligent virtual agents conference. In *Proceedings of the 18th international conference on intelligent virtual agents*, pages 17–22, 2018.

[11] D. Porfirio, E. Fisher, A. Saupp&e, A. Albarghouthi, and B. Mutlu. Bodystorming human-robot interactions. In *UIST 2019 - Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 2019.

[12] G. Smith, J. Richardson, J. D. Summers, and G. M. Mocko. Concept exploration through morphological charts: an experimental study. 2012.

[13] M. Stickdorn, M. Hormess, A. Lawrence, and J. E. Schneider. This is service design doing : applying service design thinking in the real world : a practitioner's handbook. *This is service design doing*, 2017.

[14] C. Strassmann and N. C. Krämer. A categorization of virtual agent appearances and a qualitative study on age-related user preferences. In *International Conference on intelligent virtual agents*, pages 413–422. Springer, 2017.

[15] T. Trescak, A. Bogdanovych, S. Simoff, M. Williams, and T. Sloan. Virtual dreaming: simulating everyday life of the darug people. In *International Conference on Intelligent Virtual Agents*, pages 509–512. Springer, 2016.

[16] A. Van Boeijen, J. Daalhuizen, and J. Zijlstra. *Delft Design Guide: Perspectives, Models, Approaches, Methods.* BIS Publishers, 2020.

[17] D. Wilde, A. Vallg&aarda, and O. Tomico. Embodied design ideation methods: Analysing the power of estrangement. In *Conference on Human Factors in Computing Systems - Proceedings*, volume 2017-May, 2017.

Online filters and social trust: why we should still be concerned about Filter Bubbles

Andreas Falck[1*] & Kurtis Boyer[2]
[1]Department of Special Needs Education, University of Oslo
[2]Johnson Shoyama Graduate School of Public Policy, University of Saskatchewan
* Corresponding author: andreas.falck@isp.uio.no

Abstract

Eli Pariser's (2011) notion of a "Filter Bubble" describes the effect of social media filters tuned to predict what types of online contents social media users are likely to interact with, and subsequently presenting "more of the same" in order to maximise clicks. The Filter Bubble concept originally fuelled worries that users will find themselves in positive feedback loops, becoming exposed mostly to content that they already agree with, subsequently missing out on news and information that would contradict their pre-existing views. This initial worry has subsequently been challenged, by research showing that the views and sources that social media users are exposed to are actually quite diverse. Here, we argue that the original "Filter Bubble" theory, as well as subsequent criticisms, rest on a too simplified model of human belief formation, in which information content is over-emphasised at the expense of social dynamics. We argue that filter bubbles are still problematic, as they moderate peer feedback in a way that distorts how we evaluate information together with others.

Introduction

In platform-mediated online interactions, the information flow is often restricted in a novel and worrisome way: algorithms, tailored around the platform's instrumental goal to maximize clicks and sell ads, controls both the information that reaches the user, and whom do the user's output on social media (i.e. "posts") reach. This is a source of concern, not because the information flow is restricted per se (all media constrains the flow information in some way), but because it is designed to be undetected by the user. This sets it aside from most forms of online moderation, and many instances of censorship, where the user would at least be aware of how their information channels are tampered with. While we do not know the exact weights and criteria of the content-curating algorithms of platforms like Facebook and Youtube, Eli Pariser (2011) formulates the gist of their operation:

"Internet filters looks at the things you seem to like – actual things you've done, or the things people like you like – and tries to extrapolate. They are prediction engines, constantly creating and refining a theory of who you are and what you'll do and want next."

According to this model, the algorithms predict what kind of information the user is likely to interact with, and shows this kind of information more often. On this backdrop, Pariser goes on to define the *Filter Bubble* as an information-based phenomenon:

"Together, these engines create a unique universe of information for each of us – what I've come to call a filter bubble."

In short, the Filter Bubble according to Pariser (2011) is the information landscape resulting from the operation of social media algorithms whose goal is to maximise user interaction with the network ("engagement"), in order to sell clicks and ad space. The seeming consequence is that individuals are presented with a too restricted selection of perspectives and information, so that their pre-existing ideas are reinforced in a positive feedback loop. Thus, the main problem with the Filter Bubble is supposed to be the selection of information that the social media user encounters: the users will simply not be exposed to potentially "good" ideas to a sufficient extent. This concern follows the tradition of previous research on misinformation in online and offline settings. A large empirical study by Flaxman and colleagues (2016) calls into question the very essence of the argument, as they showed that social media users are exposed to more diverse views and news than non-users. Following results like this, many authors have thus suggested that the initial worries about filter bubbles are unfounded (Zuiderween Borgesius, 2016; Bruns, 2019; Dahlgren, 2021). However, few commenters take into account the social context of belief formation, which is not so much about which information is available, but about which information to trust. Here, we will argue that it is the social dynamic of the online environment, and not the information landscape per se, that is affected negatively by click-maximizing social media algorithms. To do so, we must first discuss the social context of knowledge formation outside of social media contexts.

The Adaptive Features of Unmoderated Social Interaction

No human being would get along in their world without the aid of others. Knowledge is no exception: we rely on others to form knowledge, and culture implies that we build upon the knowledge of previous generations (Boyd & Richerson, 2009). Moreover, there exists a body of evidence that beliefs and sentiments that we share with others are privileged in human cognition. We encode information more strongly if we believe it to be attended to by others (Shteynberg, 2010). In addition, the valence of the information itself is inflated when it is shared with others: funny videos are judged as funnier (Fridlund, 1994), and persuasive political speeches are judged as more persuasive (Shteynberg et al., 2016). Research on groupthink (Turner & Pratkanis, 1998) and conformity (Baron et al. 1996) suggest that we tend to accept the beliefs that are salient in our social group.

The advantage of forming beliefs by drawing on those around us becomes apparent if we think about belief formation through the lens of evolutionary psychology. Humans have throughout history depended strongly on others within their social group for survival. While groups in pre-industrial settings were often formed incidentally around variables such as kinship or proximity, they were often kept together by instrumental goals of profit or survival, which in turn provide an external metric to judge information by. For example, if an agrarian community neglects harvesting the crops in time, the consequences could be devastating for both the group and for the individual.

In order to attain such critical goals, it is in the common interest of the group to find the best common understanding of any situation. Hugo Mercier and Dan Sperber have recently (2017) shown how social interaction promotes truth-seeking beyond what any individual can achieve. They point out that the so-called confirmation bias (Nickerson, 1998), which is counted among the heuristics that leads to choice error, apply selectively to views held by oneself. Therefore it supports correctness in social settings: arguing for one's position is optimized by selecting positive evidence, whereas others are better suited to question one's argument (Mercier & Sperber, 2017). Likewise, when peers fail to find flaws in one's arguments, then the corresponding beliefs are likely to spread in the group. Open discussion is therefore a regulatory system: the group uses positive and negative feedback to support good arguments, and pruning bad arguments, as to (paraphrasing Whitehead and Popper) let mistaken beliefs die instead of their carriers. Importantly, this happens in public discourse, not in individual minds.

When this system works as expected, the consensus of the group has a heuristic value as a guide to truth, or at least, to collective action that is effective for attaining the goal at hand. This explains why humans are more inclined to accept beliefs that are perceived as shared with their social group. As social interaction facilitates truth-seeking on the group level, then the perceived consensus among the group becomes a useful cue to truth for the individual. Negotiating beliefs in social interaction has been an adaptive strategy throughout human history. However, for this to be adaptive two conditions must be in place: First, there needs to be a free exchange of ideas, where negative feedback is allowed. Second, the individual needs to have an accurate perception of who takes part in the discourse, without which they will have a false impression of to what extent beliefs are being shared. Next, we will argue that many social media platforms pose problems for both these conditions.

How Social Media Distort the Context of Belief Evaluation

The platform's selection of information not only affects the kinds of information the user consumes, but also distorts the selection of peers that the user interacts with. Hence, the user does not only get to see more information that they tend to agree with, they will interact more with the people they tend to agree with, and less with those they disagree with. This may lead to the impression that more people share one's views than is actually the case. Not only will users see fewer social media posts contradicting their pre-existing views, they would also get less negative feedback on the views which they advertise through posts: simply because fewer of the peers that would disagree would actually see the message. Facebook's "friends list" makes a case in point here. Even if only a small subset of a particular user's Facebook friends have views similar to their own, on a specific topic, their views on this topic would be a larger part of the user's information flow. While users in principle could assess the number of friends whose views they typically hear (e.g. by comparing the posts visible on individual friends' pages with the posts appearing in the user's feed), it is unlikely to be done regularly by most users. Similarly, my posts as a Facebook user overtly appear to be broadcast to "my (Facebook) friends", while in reality they would reach these friends differentially. If human rationality relies on having our views tested against people we trust, but those whose reactions would be most valuable to this end never sees the content we post, the virtuous social feedback mentioned above is diminished. Confirmation bias is still active, but it has lost its adaptive quality suggested by Mercier and Sperber (2017). Since the actual social network is tampered with by the filters, peer feedback becomes less effective in helping users assess their own convictions. The result is a false sense of consensus, in which many of the user's pre-existing beliefs and convictions will appear as shared with the user's group, when they are actually not.

Contradicting information within our bubbles: how is it perceived by the user?

We will make a final point about the encounters social media users do experience with information that contradicts their pre-existing views. As pointed out by Flaxman and colleagues (2016), social media users are by no means isolated from views they do not agree with. However, this fact is more compatible with the Filter Bubble concept than the original information-centred view suggests. Recall that the content-curating algorithms predict what information we are likely to interact with, rather than simply what we are likely to endorse or like. Therefore it may make sense for the algorithms to select more radical and extreme views regardless of leanings, as these are expected to generate more interest from users[1]. Extreme

[1] We thank an anonymous reviewer for this insight.

views are however less likely to change someone’s mind across political boundaries, so it would not help nuancing the discourse. One may also ask how people engage with views they don’t agree with. We conjecture that among the arguments and views that one do not endorse, the main mode of engagement beyond consuming these posts, is arguing against them. However, this presupposes that one can formulate the counter-argument, i.e. one finds the opposing view weakly argued in the first place. The risk is thus, that the algorithm over time learns to present us the counter-arguments (against our views) that we already perceive as unconvincing, while becoming less likely to present those counter-arguments that have potential to change our minds. On the larger scale, users may end up with the impression that the opposing side has worse arguments than they actually have. Whether this may be the case has to our knowledge not been investigated, and ought to be targeted by future studies. Because of these considerations, the fact that contradicting views are encountered inside our bubbles, is not by itself evidence against adverse effects of social media algorithms.

Conclusion

In sum, click-maximizing social media platforms such as Facebook curate not only users’ access to views and opinions, but also the social context in which views and opinions evolve. The epistemic virtues of social interaction are attenuated, and the user’s own beliefs become inflated by how they appear to be shared with the user’s group. This warrants further caution regarding filter bubbles and related phenomena, despite the accessible (and accessed) media landscape being more diverse than ever before. More importantly, research about social media would benefit from widening its scope, to take into account the social context of human cultural evolution to a larger extent.

Acknowledgments

We thank the two anonymous reviewers for their valuable comments and insights. An earlier incarnation of this paper was presented as Boyer & Falck (2021) at the 2021 PERITIA conference “Trust in Expertise in a Changing Media Landscape”. The authors wish to thank the attendants of said conference, for many valuable comments. A.F. acknowledges support from the Swedish Research Council (grant no. 2016-06783).

References

Baron, R. S.; Vandello, J. A.; Brunsman, B. (1996). "The forgotten variable in conformity research: Impact of task importance on social influence". *Journal of Personality and Social Psychology*. **71** (5): 915–927. doi:10.1037/0022-3514.71.5.915.

Boyd, R. and Richerson, P. J. (2009) ‘Culture and the evolution of human cooperation’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1533), pp. 3281–3288. doi: 10.1098/rstb.2009.0134.

Bruns, A. (2019). Filter bubble. *Internet Policy Review*, *8*(4). https://doi.org/10.14763/2019.4.1426

Dahlgren, P. M. (2021). A critical review of filter bubbles and a comparison with selective exposure. *Nordicom Review*, *42*(1), 15–33. https://doi.org/10.2478/nor-2021-0002

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, *80*(S1), 298–320. https://doi.org/10.1093/poq/nfw006

Fridlund, A. J. (1991) ‘Sociality of solitary smiling: Potentiation by an implicit audience’, *Journal of Personality and Social Psychology*, 60(2), pp. 229–240. doi: 10.1037/0022-3514.60.2.229.

Mercier H, Sperber D (2017) *The enigma of reason*. Cambridge, MA: Harvard University Press.

Nickerson, R. S. Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* 2, 175 (1998).

Pariser, Eli. (2011). *The Filter Bubble: What the Internet is Hiding from You.* London: Penguin UK.

Shteynberg, G. (2010) ‘A silent emergence of culture: The social tuning effect.’, *Journal of Personality and Social Psychology*, 99(4), pp. 683–689. doi: 10.1037/a0019573.

Shteynberg, G. *et al.* (2016) ‘The broadcast of shared attention and its impact on political persuasion.’, *Journal of Personality and Social Psychology*, 111(5), pp. 665–673. doi:
10.1037/pspa0000065.

Turner, M. E.; Pratkanis, A. R. (1998). "Twenty-five years of groupthink theory and research: lessons from the evaluation of a theory". *Organizational Behavior and Human Decision Processes*. **73** (2–3): 105–115.

Zuiderveen Borgesius, F. J., Trilling, D., Möller, J., Bodó, B., de Vreese, C. H., & Helberger, N. (2016). Should we worry about filter bubbles? *Internet Policy Review*, *5*(1). https://doi.org/10.14763/2016.1.401

The Crisis of Trust in AI and Autonomous Systems

Amandus Krantz[1]

[1]*Lund University Cognitive Science*

amandus.krantz@lucs.lu.se

The future is robotic. Already we are seeing how society is changing with self-driving cars and robots at hospitals and schools. The considerable potential of autonomous systems (AS) is highly discussed. What is not highly discussed, and rarely even acknowledged, is the key role trust plays in realizing these benefits and the problems this may cause for human-AS interaction research. There currently exists no common way of defining, testing, or measuring trust. This lack of common foundation, both for trust in general but also in human-AS relations, may at best result in sporadic progress and adoption of these systems and may at worst lead to public disillusionment and abandonment, delaying the potential benefits of AS.

According to Glikson & Woolley (2020) the match between a user's trust in technology and that technology's abilities is a predictor for future use; low trust in capable technology leads to disuse, while high trust in incapable technology leads to frustration which leads to misuse which in turn may lead to dangerous situations. For example, a user with low trust in the capabilities of their robotic vacuum cleaner will be more inclined to vacuum manually, negating the benefits of the robotic vacuum. On the other hand, a user who puts too much trust in their self-driving car's ability to avoid obstacles may feel comfortable enough to sleep at the wheel, potentially causing accidents if the car encounters an obstacle it cannot avoid. Enholm, Papagiannidis, Mikalef, & Krogstie (2021) agrees that for an AI-system to be used at all in a business setting, the user must have some level of trust in it. Marsh (1994), in an early attempt at creating a taxonomy for trust in human-AS interaction, writes that trust should be considered a fundamental part of cooperation and communication.

Given this, it would seem that research on trust, both in general and in human-AS relations, should be a highly prioritized area. Understanding how trust works could reduce resources wasted on disused technology, and minimize the dangers that come with misuse and over-reliance on incapable technology, making trust research beneficial for society. Unfortunately, this is often not the case. Research that focus on how trust works and how to measure it in human-AS relations is pretty limited, and the little focused research that does exist is often plagued by several problems.

The first of these problems is the use of short, single measure, experiments. Trust is not a constant, it changes as the interaction proceeds, it is dynamic (Blomqvist, 1997). A single question about trust at the end of a study only shows what the participant thought at that particular time, but tells you nothing about how the experiment actually impacted the trust (Glikson & Woolley, 2020).

Second is the problem of unclear terminology (Cameron et al., 2021; Jessup, Schneider, Alarcon, Ryan, & Capiola, 2019). Trust in AS is typically presented in terms of performance and reliability; however, there is a second type of trust that is more general, established before one can make a rational judgement about reliability. It is based more on instinct, emotions, and gut feeling (Fiske, Cuddy, & Glick, 2007; Marsh, 1994; McAllister, 1995). Without clear terminology to indicate which type of trust is being measured, researchers run the risk of measuring something they are not intending (Chita-Tegmark, Law, Rabb, & Scheutz, 2021).

Related to the problem of unclear terminology is finally the problem of overly simple, varied, and non-standard methodology (Glikson & Woolley, 2020). There exists no common method of measuring trust (Chita-Tegmark et al., 2021; Gao, Sibirtseva, Castellano, & Kragic, 2019), leading many researchers to make use of vague Likert scale questioning, for example "On a 5 point scale, how much do you trust this robot/person/agent?". These types of questions are problematic since small changes in the scale (e.g., a 7 point scale instead of a 5 point scale) or, as mentioned, the terminology, can make it difficult, if not impossible, to generalize and compare the results with other studies (Chita-Tegmark et al., 2021). Using home-brew methodologies may also cause issues with statistical significance, as shown by Schrum, Johnson, Ghuy, & Gombolay (2020) who discovered that only 3 of the 110 peer-reviewed human-robot interaction papers they examined had properly implemented and analysed their questionnaires.

Some attempts have been made to create methodologies for the measurement of trust (Bartneck, Kulić, Croft, & Zoghbi, 2009; Berg, Dickhaut, & McCabe, 1995; Schaefer, 2016). They have, however, failed to reach any kind of common usage as there seems to exist some doubts about whether they are transferable across field boundaries (Glikson & Woolley, 2020). For example, the investment-style games proposed by Berg, Dickhaut, & McCabe (1995) seems to work well for human-human trust related to investments, but the methodology may not work as well for human-AS interaction as it may require the participant to make unrealistic assumptions about the capabilities of the AS (e.g. its intelligence or level of autonomy) which may impact the reported level of trust.

At the heart of these problems lies the more in-depth problem of defining trust. No commonly accepted definition of trust currently exists, and progress towards creating one is slow. Blomqvist (1997), giving an overview of the many definitions of trust, writes that of the fields investigated, only the field of social psychology has a reasonable definition of trust, while moral philosophy and economics either do not address the topic at all or have created definitions that allow them to ignore it. Yet, the word trust is used in pretty much every field, from AI to political science to law. It is used so much, in so many fields, that one almost has to assume that it is referring to the same concept. However, trust in a human and trust in technology are two very different things, and trust in technology and trust in AS is another one still (Glikson & Woolley, 2020). Distinctions like these are vitally important when developing methodologies and measures for trust studies, as a methodology that works for evaluating trust in humans may be nonsensical when applied to trust in a self-driving car or humanoid robot. Researchers have to keep this in mind not only when transferring measures and methodologies from human-human trust research to human-AS research, but also when dealing with different types of AS. Trust in a self-driving car, for example, may not work the same as trust in a robotic vacuum cleaner, factory robot, or military drone.

Trust, then, should be considered a fundamental part of human-AS interaction, and is likely a requirement for cooperation between humans and AS to even start (Enholm et al., 2021; Marsh, 1994). Yet, our understanding of what it actually is, how it behaves, and its mechanics is very limited, and attempts at increasing this understanding are often hindered by fundamental problems, such as unclear terminology and methodologies that make results difficult to generalize.

A fully unified and universal definition of trust may not be possible, but we must at least attempt, through interdisciplinary efforts, to find a common foundation from which discussion and progress can grow. If we do something about this fundamental problem at this early stage by establishing a solid foundation, well implemented methodologies, and clear terminology, we will have the chance to gain an increased understanding for one of the most fundamental requirements for society, communication, and interaction.

Acknowledgement

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

References

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, *1*(1), 71–81. https://doi.org/10.1007/s12369-008-0001-3

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Blomqvist, K. (1997). The many faces of trust. *Scandinavian Journal of Management*, *13*(3), 271–286. https://doi.org/10.1016/S0956-5221(97)84644-1

Cameron, D., de Saille, S., Collins, E. C., Aitken, J. M., Cheung, H., Chua, A., … Law, J. (2021). The effect of social-cognitive recovery strategies on likability, capability and trust in social robots. *Computers in Human Behavior*, *114*. https://doi.org/10.1016/j.chb.2020.106561

Chita-Tegmark, M., Law, T., Rabb, N., & Scheutz, M. (2021). Can You Trust Your Trust Measure? *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 92–100. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3434073.3444677

Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2021). Artificial Intelligence and Business Value: A Literature Review. *Information Systems Frontiers*. https://doi.org/10.1007/s10796-021-10186-w

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Gao, Y., Sibirtseva, E., Castellano, G., & Kragic, D. (2019). Fast adaptation with meta-reinforcement learning for trust modelling in human-robot interaction. *ArXiv:1908.04087 [Cs]*. Retrieved from http://arxiv.org/abs/1908.04087

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, augmented and mixed reality. Applications and case studies* (pp. 476–489). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-21565-1_32

Marsh, S. P. (1994). Formalising trust as a computational concept (PhD). University of Sterling.

McAllister, D. J. (1995). Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Academy of Management Journal*, *38*(1), 24–59. https://doi.org/10.5465/256727

Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the "trust perception scale-HRI". In R. Mittu, D. Sofge, A. Wagner, & W. F. Lawless (Eds.), *Robust intelligence and trust in autonomous systems* (pp. 191–218). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4899-7668-0_10

Schrum, M. L., Johnson, M., Ghuy, M., & Gombolay, M. C. (2020). Four years in review: Statistical practices of likert scales in human-robot interaction studies. *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 43–52. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3371382.3380739

Artefactual ethics as opportunity for rethinking “natural” ethics

Joel Parthemore[*]& Blay Whitby[†]

Abstract

This paper serves as introduction to a significantly longer paper in progress. It argues that, within the ethics community, the wider philosophical establishment and society in general, people have been far too lax about what to accept as morally “right” behaviour – far too quick to let themselves and, all too often, each other off the hook. By drawing comparisons to artefactual behaviour and the objections people raise to calling that behaviour the morally acceptable behaviour of authentic moral agents, this paper lays out a framework by which human ethics and meta-ethics can more fruitfully be approached. An earlier paper of ours (Parthemore and Whitby, 2014) argued that, for an action to be morally right, one must have a convergence of the right motivations, the right means, and the right consequences. The underlying insight is that deontological, virtue-ethics-based, and consequentialist accounts all have their necessary role to play, but each tends to get too focused on itself and its merits to the loss of the bigger picture; while utilitarian accounts, as perhaps the most prominent division within consequentialism, face the further problem of failing to allow for those occasions where the needs of the few, or the one, outweigh the needs of the many, as Ursula K. LeGuin (1973) so devastatingly addressed. Although the requirement to align motivations, means, and consequences may seem impossibly onerous, it need not be, provided one is prepared to allow that moral behaviour is far more difficult to achieve, either for artefacts or human beings, than it might seem at first glance. Mistakes will be made. Perhaps it matters more to take responsibility for those mistakes than to assure oneself, despite reasonable argument to the contrary, that one has avoided them. It is time to hold artefactual and natural agent alike to a higher standard.

1 Introduction: Human beings, artefactual agents, and the responsibility game

For purposes of this paper, we will take moral agency as the capacity to take responsibility, and be held responsible, for one’s actions.[1] Intimately wrapped up in all matters moral is the responsibility question: who individually has, and who collectively have, responsibility for any given action or set of events.[2] People have been attributing all manner of agency to virtual and physical artefacts – including moral agency – at least since the advent of Eliza. With the advent of “self-driving” cars and “autonomous” battlefield robots (see, e.g., Sharkey 2011a,b), the responsibility question has only grown. In attempting to answer it, researchers interested in artefactual moral agency (e.g., Wallach et al., 2011; Allen et al., 2000) have tended to focus more-or-less equally on what artefacts do and what they fail to do – what morally relevant “choices” they make or fail to make – and here the standard objection is that existing artefacts either do the “wrong” things (e.g., battlefield robots producing “friendly fire”) or fail to do the “right” ones (say, making no response on seeing someone in danger, as with the Uber car that failed to brake for the pedestrian in Arizona). Setting aside whether it provides an adequate litmus test for moral agency – as it is surely attempting to do – Colin Allen and colleagues’ (2000) proposed Moral Turing Test[3] sets a standard that, it would seem, no existing artefact could pass. Over-attribution of moral agency is, seemingly, met by bald under-performance.

[*] Adjunct researcher, University of Skövde, Sweden; joel.parthemore@his.se

[†] Visiting lecturer, University of Sussex, UK; B.R.Whitby@sussex.ac.uk

[1] This is to make the traditional distinction from *moral patienthood,* which may usefully be described as an entity having certain moral responsibilities attached to it on the part of moral agents (see, e.g., Pluhar, 1988).

[2] Needless to say, neither individual nor collective responsibility excludes the other.

[3] In brief, a purported agent is a moral agent if it takes what they consider the morally “right” decision a sufficiently high percentage of the time.

1.1 Action, inaction, and intention

Somewhat by contrast, psychology tells us that people are, *ceteris paribus,* far more willing to excuse inaction in themselves or others – a failure to act – than to excuse actions they consider morally problematic.[4] To fail to save someone's life – to allow that death to happen – is generally considered less morally wrong than to take a life, even if the two circumstances are, in all other relevant aspects, the same. At the same time, it seems difficult how one might logically justify how the passive vs. active nature of the behaviour could make the necessary difference – as, e.g., Sisela Bok (1999) has pointed out in discussing the nature of lies. How is a *lie of omission* (what I fail to tell you) any less a lie than a *lie of commission* (what I tell you wrongly)? If the one is morally problematic, then so is the other.

Along similar lines, the *Doctrine of Double Effect* (DDE) – often invoked to uphold Roman Catholic thinking on abortion – holds that knowing that something otherwise morally unacceptable will happen as the unintended consequence of one's actions (or inactions) is at least sometimes acceptable whereas intending that same thing to happen would not. The doctrine is necessary for reconciling moral absolutes (killing of human beings is always wrong; human foetuses are human beings; therefore abortion is always wrong) with real-world cases that would otherwise pose problems for those moral absolutes. (What if allowing the pregnancy to go to term – not performing an abortion – would kill the mother or both the mother and the child? Many defenders of the DDE would argue that that is morally preferable because the death of the child, though foreseeable and unfortunate, is not intended; whereas abortion is always an intentional act.) The trolley problem, as originally formulated – quite succinctly![5] – raises difficulties here, as the DDE can equally be used to argue for saving the life of the one person on the one track (with the unintended consequence of killing five on the other) or for saving the lives of the five at the cost of the one: it all depends on one's intentions, which Foot (rightly, we believe) declares unacceptable.[6] For Foot, intention is important but insufficient; means matter; and, clearly, she takes a utilitarian-inspired interest in numbers in favouring the lives of the five over that of the one. For Foot, the outcome *must* be weighed along with the means and intention. For all her sympathy with those who oppose abortion and support the DDE, she sees merit not only in saving the mother's life at the deliberate loss of the child's – i.e., via abortion – when both would otherwise be certain to die; but also in pursuing abortion in cases where only one or the other might be saved. Foot rescues a version of the DDE at the loss of the possibility of absolute moral principles; but one might see this as a good thing. Claims to absolute moral principles may serve to excuse behaviour, as that by persons inclined to take a dogmatic position on abortion, that perhaps should not be excused. If artefacts are not allowed resort to sophistry – whether we think them capable of genuine sophistry or not – then neither should people be.

1.2 Hard-and-fast rules, rules of thumb, and ground rules

Much ink has been spilled within the machine ethics community on what rules to hardwire into artefactual moral agents, and much effort has been made to draw inspiration from Isaac Asimov's *Three Laws of Robotics* – despite the many times, in his stories, where Asimov showed just what impossible conundrums those rules created: a rule intended to anticipate every possible circumstance rarely if ever can. Such rules set a bar so high that not even those who clearly qualify as moral agents can reach it, never mind those whose moral agency may be considered in dispute. If artefacts are ever to be considered candidates for moral agency, then they should be held to no higher a standard than what human beings can achieve.

Rules of thumb might fare better. First-order predicate logic may rely on universal quantification, but the *lifeworld* (Husserl, 1970)with which people engage on a daily basis has a habit of throwing up exceptions. That said, if Foot is right – and we think she is – then *any* strictly rule-based approach will fail. Perhaps the lesson to be learned from present-day artefacts, and the reason so few are willing to grant them moral

[4] If one objects that no one could excuse the human equivalent of the Uber case, the authors have personally encountered it more than once.

[5] "...It may be supposed that [the man] is the driver of a runaway tram which he can steer from one narrow track onto another; five men are working on one track and one man on the other; anyone on the track he enters is bound to be killed" (Foot, 1967).

[6] "A certain event may be desired under one of its descriptions, unwanted under another, but we cannot treat these as two different events, one of which is aimed at and the other not. And even if it can be argued that there are here two different events... the two are obviously much too close for an application of the doctrine of double effect" (Foot, 1967).

agency,[7] is not that they lack the right rules with which to make the right decisions; rather it is that they lack the capacity to make decisions or take responsibility for them in the first place – in any but the most loosely metaphorical of senses. Remember that, by our definition, moral agency requires the capacity to take responsibility: something that – in company with newborn infants and certain among the mentally infirm[8] – present artefacts would appear to lack. Most infants and at least some mentally infirm persons can be expected to outgrow their present conditions; by contrast, no amount of time and patience will change present-day artefacts or their close kin into moral agents.

This is not to say that one can or should avoid hard-and-fast rules altogether. At least at first blush, the principle that what is acknowledged as morally wrong should never simultaneously be accepted as morally right seems like a suitable candidate. Indeed, if one accepts that moral right and wrong are mutually exclusive, then it follows of logical necessity. Yet "lesser of two evils" arguments, widely used, require that the "lesser" evil is, at the least, morally acceptable if not strictly speaking "right"; and "just war" accounts – to take one example – critically depend on such arguments. The evil action (or inaction) becomes the good because, it is said, there is no alternative. Jean-Paul Sartre showed that, on nearly every occasion where people claim a lack of alternatives, there *are* alternatives; the problem is either that we fail to see or that we fail to acknowledge them. If people would not accept "lesser of two evils" arguments to excuse artefactual behaviour – and we believe that few would – then they likewise should not accept them to excuse their own.

The solution posed by the full paper is to let go of moral absolutes – few things indeed are *always* morally right or wrong – and to embrace personal responsibility, as Sartre (1946) has challenged us all to do: taking responsibility and acknowledging both when we believe that we *have* done right, despite all evidence and arguments to the contrary, with a willingness and ability to defend the reasoning that led us there; and when we know we have done wrong, either because we could not see an alternative or lacked the courage to embrace it. The proper response to the high standards imposed on moral agency for artefacts is not to lower those standards on artefacts but use them to raise the bar for ourselves.

Section Two of the intended full paper, currently a work in progress, will examine the ethical theory that serves as the foundation for this extended abstract – one that calls for a convergence of the "right" motivations, the "right" means, and the "right" outcomes – and consider how it can be made to work.[9] Section Three will consider the consequences of applying that framework to purported artefactual agents. Section Four will offer three case studies: one from the field of autonomous vehicles, one from aviation, and one from medicine. Section Five will bring the artefactual lessons back to the human case and offer prescriptions on the way forward.

References

Allen, C., Varner, G., and Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3):251–261.

Bok, S. (1999). *Lying: Moral choice in public and private life.* Vintage. First published 1978.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, 5:5–15. Available online from https://philpapers.org/archive/footpo-2.pdf (accessed 26 January 2020).

Husserl, E. (1970). *The Crisis of European Sciences and Transcendental Phenomenology: An Introduction to Phenomenological Philosophy.* Northwestern University Press. tr. David Carr. First published (in German) 1954.

LeGuin, U. K. (1973). The ones who walk away from omelas. In Silverberg, R., editor, *New Dimensions*, volume 3, pages 1–8. Doubleday.

Parthemore, J. and Whitby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness*, 6(2):1–21.

[7] . . . Despite the haste with which others would do so!

[8] . . . Who nevertheless qualify as moral patients!

[9] A reviewer suggested basing that discussion around climate ethics but that, to our mind, would be a different paper. For better or worse, we have chosen to return to the artefactual question we addressed in our earlier papers.

Pluhar, E. (1988). Moral agents and moral patients. *Between the Species*, 4(1):32–45.

Sartre, J.-P. (1946). The flies. In *The Flies and In Camera*. Hamish Hamilton. tr. Stuart Gilbert.

Sharkey, N. (2011a). Automating warfare: Lessons learned from the drones. *Journal of Law Information and Scienc*, 21:140–154.

Sharkey, N. (2011b). Killing made easy: From joysticks to politics. In Lin, P., Abney, K., and Bekey, G. A., editors, *Robot Ethics: The Ethical and Social Implications of Robotics*, chapter 7, pages 111–128. MIT Press.

Wallach, W., Allen, C., and Franklin, S. (2011). Consciousness and ethics: Artificially conscious moral agents. *International Journal of Machine Consciousness*, 3(1):177–192.