Embodied Conversational Agents: Trust, Deception and a the Suspension of Disbelief

ANONYMOUS AUTHOR(S)

Building trust is often cited as important for the success of a service or application. When part of the system is an embodied conversational agent (ECA), the design of the ECA has an impact on a user's trust. In this paper we discuss whether designing an ECA for trust also means designing an ECA to give a false impression of sentience, whether such an implicit deception can undermine a sense of trust, and the impact such a design process may have on a vulnerable user group, in this case users living with dementia. We conclude by arguing that current trust metrics ignore the importance of a willing suspension of disbelief and its role in social computing.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; • **Applied computing** → *Health informatics*; • **Computing methodologies** → Philosophical/theoretical foundations of artificial intelligence.

Additional Key Words and Phrases: social agents, trust, deception, dementia

ACM Reference Format:

Anonymous Author(s). 2023. Embodied Conversational Agents: Trust, Deception and a the Suspension of Disbelief. 1, 1 (March 2023), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Reminiscing is a positive process for individuals living with Alzheimer's Disease(AD), reinforcing remaining cognitive capabilities, building positive social experiences and improving quality of life. In the AMPER project our aim is to develop an App with a storytelling agent that supports reminiscing between a user and their carer based on autobiographical memory [14]. Autobiographical life stories are important not only for human social interactions but for engagement and bonding with artificial agents [3, 6]. Such agents must be equipped with *human-like* memory processes to meet user expectations of life-likeness, intelligence and responsiveness [4, 11, 12, 15].

Reminiscence is more than mere storage and retrieval of facts, encompassing emotion, selection, association, reconstruction and adaptation to context [2, 5]. AMPER will enable a model of reminiscence by addressing these strong contextual factors. In particular, we will explore the interaction between memory and life stories by way of an embodied conversational agent (ECA) with a novel biologically-inspired autobiographical memory that can generate natural and believable life stories based on social context.

AMPER focuses on co-creation, a user-centred collaborative innovation approach where project members and stakeholders are equal partners. This enables recognition of fresh perspectives and provide transparency for both parties on what the system is capable of. The agent will perform carer-assisted intervention through storytelling by retrieving past memory of AD individuals from respective generational and personalized repository so that the narrative is meaningful and directly relevant to the individual. It will do so in a biologically plausible way applying memory

Manuscript submitted to ACM

Manuscript submitted to ACM

algorithm such as spreading activation to relate different elements of the stories so that the story appears coherent and natural.

However, the design of the system raises a complex question regarding trust and deception. Bickmore et al. point out, *"many researchers feel that they are somehow crossing an ethical boundary if their agents start discussing their childhood home or the fight they just had with their (presumably human) spouse"* [3]. However, their work also shows that giving an agent a backstory leads to greater user engagement rather than causing rejection because of the perception of being deceived.

Although in AMPER we are not planning to generate artificial back stories we do intend the agent to ask open ended questions, give positive feedback to increase confidence and to encourage further dialogue. Implicitly by meeting user's expectations of life-likeness, intelligence and responsiveness we will aim to design *human-like* appearance, characteristics, and responses into the ECA. Given we are dealing with vulnerable users and potentially asking for the disclosure of cherished lifetime memories. This raises some key questions:

- Are we deceiving people into developing a trusting relationship with an artificial system?
- Should we use such a system to encourage a user to disclose potentially sensitive information?
- Should we take into account the vulnerable nature of the users?
- If this is a deception does it matter if it achieves a positive outcome?

2 DECEPTION OR ENGAGEMENT

Bickmore et al. [3] concludes that the extent of deception in creating a *human-like* agent is unclear, users are may not be upset by it and it succeeds in the aim producing an effective design. In contrast Fisk [8] in his address to the Ethics Forum at IROS 2022 takes a more serious view of the process of giving a deceptive sense of sentience to systems such as robot companions.

"There are, as a consequence, dangers if we promote robots (in either physical or virtual environments) on a basis that they can be companions or substitute carers - with which there can be 'meaningful' interactions. Robots cannot have emotions or feelings, they can only imitate. They are machines, human beings are not. Any imagined robot 'companionship' in the mind of a person who is 'cared' for can, therefore, only be founded on deception."

It is important to emphasize here that Fisk is discussing robot companions for older individuals. Whereas AMPER's aims are closer to Bickmore et al. [3] in focusing more on specific applications and systems to address specific needs. Artificial companions in contrast are offering a system which builds trust and an emotional relationship with a user as a general requirement whether or not specific services it wishes to offer need it or not. Whether this would replace or extend human contact, how a user's data is used and owned and who is ultimately responsible for the systems actions are serious ethical considerations with Fisk arguing such systems are dystopian not symbiotic. Aylett et al. [1] supports these concerns by arguing that the *companionship* model is driven from an out-of-date commercial model that indeed seeks to foster dependence and realize a means of monetizing a user's data without any direct responsibility for the system that is deployed.

3 SOCIAL ACTORS VS DEPICTIONS OF SOCIAL ACTORS

To help untangle some of these ethical issues it is useful to return to the notion of computers as social actors as promoted by Nass (e.g [13]). This was put forward nearly two decades ago and since then we have seen a revolution in social computing with both positive and negative results. We are now able to build ECAs that, if only for short time, are

indistinguishable from the film footage of a real person. If anything, the development of ECA technology has encouraged the view of these systems being *human-like* and demonstrating *intelligent-like behavior*.

But as Clark points out, in reality such systems are not social actors, they are created to be seen as such, they are in reality depictions of social actors. *"With social robots, people use the depiction proper to engage in the pretense that they are interacting with the agents depicted. People distinguish these agents from the authorities responsible for them."* [7]. Regarding ECAs as depictions does not mean they lack power, or that they cannot be used to deceive people. Rather it places them together with other fictions that we are more familiar with: actors in a theater play, stories in a work of fiction, events in a non-documentary film.

All these cultural forms demonstrate the human ability to suspend disbelief and treat something as *real* when they know it is not. If we look at these forms we also see a similar fear of deception. A film can be propaganda or present ideas that people find unacceptable or dangerous but it is also art and can have a profound and positive effect. Story-telling is age old cultural behavior and its relationship with truth has always been problematic. Context is crucial to all these art forms. By taking into account the context you can decide if a film is propaganda or a story-teller is an entertainer or someone delusional. For example, asking an actor to act in a play is different from asking them to pretend to be a close friend at a real social gathering.

Regarding ECAs as depictions of social agents complicates what we mean by having trust in it. The questions we ask a user of a flight booking system are not the same questions we might ask someone who is watching a film. The trust we develop for artists is about the type, consistency, appropriateness and enjoyment that a a depiction brings. We know much content in a fictionalized account of a historical event are not true in the way we regard historical truth, but we may trust the artist to depict the event in a way which we find entertaining and offering a different sense of truth.

4 MEASURING TRUST WITH ECAS

As part of our co-creation work on AMPER we engaged a set of stakeholders to help guide the design of the ECA we were proposing. The issue of trust was highlighted as a potentially critical requirement: for example a positive appearance for the agent would be *"knowledgeable, trustworthy, friendly, storytelling."*, that the immediate reaction to the characters appearance needs to be *"friendly amicable trusting"*, even that the agent might have the appearance of a real person that was a trusted source such as David Attenborough.

Thus, within our project we have two types of trust that we wish to engender, the first is in the ECA, the second is in the overall application. Questionnaires that have been developed to measure user trust in computer systems via self-reporting (e.g [9, 10]) focus on areas such as deception, confidence, dependability, integrity and security. While these may be appropriate for the App as a whole, with regards to responding to the agent it leaves out the possibility of a willing suspension of disbelief.

For our project, where we are dealing both with vulnerable users and their carers, there is an important question of whether they do believe that the ECA has real sentience, or whether they are happy to pretend it has. This has an important impact on perceived trust, especially over a longer period of time when we may see a shift between these two perspectives. The ability to pretend is key to many positive social interactions and also for individuals to play out conflicts and ideas alone. We do not regard ourselves as deceiving a child when we give it a soft toy, and we are not aghast if the child speaks to it and pretends it is a social actor. Making use of this fundamental human trait, of believing and not-believing at the same time, is a valuable asset for the design of ECAs. As Bickmore et al. [3] discovered, using it can make a big difference to how effective a system is for a specific goal. From a purely practical perspective incorporating this sense of make-believe into trust questionnaires would be a useful next step.

5 CONCLUSION

We should not be complacent about the concerns Fisk raises with using robots as social companions. However, if we see the deception we wish to use in our design as similar to the deception in a work of fiction (one licensed by context) we can see how this is both acceptable as well as beneficial in terms of user experience. Looking at this problem in the context of a real project and application is a useful process and we expect to return to the trust literature in detail to help support our forward design goals.

In terms of our key questions what do we propose?

Are we deceiving people into developing a trusting relationship with an artificial system?

Possibly but not necessarily. As we have argued it is possible for a user to have a trusting relationship with and artificial system they know is a deception. However, as we test our system it is important to ascertain where belief and suspension of disbelief lies.

Should we use such a system to encourage a user to disclose potentially sensitive information?

No. But we must be aware that sensitive information may be disclosed and must be treated ethically and responsibly.

Should we take into account the vulnerable nature of the users?

Absolutely. Belief, suspension of disbelief and the impact it may have may be significantly different for these user groups.

If this is a deception does it matter if it achieves a positive outcome?

Yes. But it is important to stress this not deception in terms of the App where study purpose and information is provided to the users/AD individuals/carers who are also co-creators of the project. Rather, it is a deception in terms of creating a *human-like* fiction which may be acceptable and even desirable. AMPER follows the BPS Code of Human Research Ethics[1]. However, the BPS code deals with deception in a more concrete manner (giving false information in order to perform an experiment) rather than the deception involved in building a system that gives the impression it is alive when it is not. By avoiding techniques like creating backstories and avoiding objectives like offering companionship, AMPER avoids many of the serious problems associated with ECAs and vulnerable users. However, it does matter that we are using such a fiction and it will be important to monitor it's effect on trust as well as monitor the impact it has on our users.

REFERENCES

[1] M.P. Aylett, R. Gomez, E. Sandry, and S. Sabanovic. 2023. Unsocial Robots: How Western Culture Dooms Consumer Social Robots to a Society of One. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*.

[2] Frederic Charles Bartlett and Frederic C Bartlett. 1995. *Remembering: A study in experimental and social psychology*. Cambridge university press.

[3] Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2009. Engagement vs. deceit: Virtual humans with human autobiographies. In *Intelligent Virtual Agents: 9th International Conference, IVA 2009 Amsterdam, The Netherlands, September 14-16, 2009 Proceedings 9*. Springer, 6–19.

[4] Cyril Brom, Jiří Lukavskỳ, and Rudolf Kadlec. 2010. Episodic memory for human-like agents and human-like agents for episodic memory. *International Journal of Machine Consciousness* 2, 02 (2010), 227–244.

[5] Robert N Butler. 1963. The life review: An interpretation of reminiscence in the aged. *Psychiatry* 26, 1 (1963), 65–76.

[6] J Campos. 2010. *MAY: my Memories Are Yours. An interactive companion that saves the users memories*. Ph. D. Dissertation. Master thesis, Instituto Superior Técnico.

[7] Herbert H Clark and Kerstin Fischer. 2022. Social robots as depictions of social agents. *Behavioral and Brain Sciences* (2022), 1–33.

[8] Malcolm Fisk. 2022. AI, Limitations and Illusions - Towards a Symbiotic or Dystopic Society? https://iros2022.org/program/special-forum/ethics-forum/

[9] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[1] https://www.bps.org.uk/guideline/bps-code-human-research-ethics

[10] Moritz Körber. 2019. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*. Springer, 13–30.

[11] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social robots for long-term interaction: a survey. *International Journal of Social Robotics* 5 (2013), 291–308.

[12] Mei Yii Lim. 2012. Memory models for intelligent social companions. *Human-Computer Interaction: The Agency Perspective* (2012), 241–262.

[13] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[14] Katherine Nelson. 1993. The psychological and social origins of autobiographical memory. *Psychological science* 4, 1 (1993), 7–14.

[15] Caroline Rizzi, Colin G Johnson, Fabio Fabris, and Patricia A Vargas. 2017. A situation-aware fear learning (safel) model for robots. *Neurocomputing* 221 (2017), 32–47.