UPPSALA
UNIVERSITET

UPTEC STS 19022
Examensarbete 30 hp
Juni 2019

Identifying & Evaluating System Components for Cognitive Trust in AI-Automated Service Encounters

Trusting a Study- & Vocational Chatbot

Joakim Eklund
Fred Isaksson

UPPSALA
UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Identifying & Evaluating System Components for Cognitive Trust in AI-Automated Service Encounters

Joakim Eklund & Fred Isaksson

The intensifying idea that AI soon will be a part of our everyday life allows for dreams about the complex relationship we one day could have with non-biological social intelligence. However, establishing societal and individual acceptance of AIpowered autonomy in disciplines built upon to the reliance to human competence raises a number of pressing challenges. One of them being, what system components will engender respectively counteract cognitive trust in socially oriented AI-automated processes?

This masters thesis tackles the seemingly ambiguous concept of trust in automation by identifying and evaluating system components that affect trust in a confined and contextualised setting. Practically, we design, construct and test an AI-powered chatbot, Ava, that contains socially oriented questions and feedback about study- and vocational guidance. Through a comparative study of different system versions, including both quantitative and qualitative data, we contribute to the framework for identifying and evaluating human trust in AI-Automated service encounters. We show how targeted alterations to design choices constituting the system components transparency, unbiasses and system performance, identified to affect trust, has consequences on the perception of the cognitive trust concepts integrity, benevolence and ability. Our results display a way of conduct for practitioners looking to prioritise and develop trustworthy autonomy. More specifically, we account for how cognitive trust is decreased when system opacity is increased. Moreover, we display even more concerning effects on trust due to micking contextual bias in the conversation agent

Handledare: Maria Mattsson Mähl
Ämnesgranskare: Kristiaan Pelckmans
Examinator: Elísabet Andrésdóttir
ISSN: 1650-8319, UPTEC STS 19022

Sammanfattning

Automatisering - att utföra arbete utan mänsklig inblandning, kan vara en viktig del i att effektivisera olika processer. Automatisering, som vi känner till det idag, påbörjade sin utveckling mellan den första och tredje industriella revolutionen och har sedan dess helt förändrat spelplanen för alla produktproducerande industrier. Då digitaliseringen mer eller mindre snart underligger alla samhällsprocesser, drar även tjänsteleverantörer i större utsträckning nytta av automatiserade processer. Vanligtvis berör automatiseringen av tjänster möjligheter till att t.ex. erbjuda detaljhandel eller kundsupport utan någon mänsklig interaktion. För mer socialt orienterade tjänster där mänsklig interaktion och dialog anses vara en stor del av värdet (handledning, psykiatri, studie- och yrkesrådgivning etc.) har datorer ännu inte kunnat konkurrera med den tillit vi placerar hos en mänskligt övervakad process. Artificiell Intelligens (AI), skapandet av intelligenta och självtänkande maskiner, öppnar dock upp för nya möjligheter att efterlikna och potentiellt ersätta mänskliga beteenden och intelligens. AI används redan inom flera områden av vår vardag och dess tillämpning förväntas bara att intensifieras. Övergången från konversationer mellan ansikte och ansikte till ansikte och AI i sociala dialoger kräver kunskap om hur vi ska översätta mänsklig trovärdighet till binära siffror. Att etablera förtroende kommer att vara avgörande för att säkerställa den mänskliga acceptansen och utvecklingen av AI. Särskilt i tjänster där det mänskliga interaktionsvärdet traditionellt anses vara högt.

Detta examensarbete tacklar det till synes komplexa och flyktiga konceptet tillförlitlighet i automatiserade processer genom att identifiera och utvärdera systemkomponenter som påverkar förtroende i en begränsad och kontextualiserad miljö. I praktiken designas, konstrueras och testas en AI-driven chatbot, Ava, som för en socialt orienterade konversation om studie- och yrkesvägledning. Genom jämförande studier av olika systemversioner, innehållande både kvantitativa och kvalitativa data, bidrar studien till ramverket för att identifiera och utvärdera mänskligt förtroende för automatiserade tjänstebemötanden. Studien visar hur systematiska förändringar av designval som utgör systemkomponenterna transparens, kontextuell subjektivitet och systemprestation, som identifierats att påverka tillförlitlighet, har konsekvenser för uppfattningen av de kognitiva förtroendekoncepten integritet, välvilja och förmåga. Resultaten visar ett tillvägagångssätt för utövare som vill prioritera och utveckla tillförlitlig autonomi. Mer specifikt redogörs det i studien för hur kognitivt förtroende minskar när systemets transparens minskar. Vidare, exemplifierar vi ännu mer alarmerande effekter på förtroende genom att imitera kontextuell subjektivitet i konversationsagenten.

Acknowledgement

The master's project has been performed in collaboration with the company AlphaCE Coaching & Education located in Uppsala during the spring semester in 2019. Other companies or organisations providing help along the way and are worth mentioning is; Arbetsförmedlingen, JobTechdev, Lundellska Skolan and Rosendalsgymnasiet.

First, we would like to express our sincerest appreciation to Maria Mattsson Mähl, Senior Advisor, Partner and Board member at AlphaCE for being hugely inspirational whilst supervising this master's thesis. We would also like to thank AlphaCE and their fantastic staff for housing us and always being present to answer all of our work-market related questions. Furthermore, we would like to thank Kristiaan Pelckmans at the Department of Information Technologies, Uppsala University. Thank you for your assistance and advice throughout this thesis, steering the project in the right path. Finally, for a lot of help and inspiration along the way, we would like to thank Davide Vega D'aurelio at the Department of Information Technologies, Uppsala University.

Joakim Eklund & Fred Isaksson
Uppsala, June 2019.

Distribution of Work

This master's thesis is the last and final step in the master programme in Sociotechnical Systems Engineering at Uppsala University during the spring semester of 2019. The study has been conducted by Joakim Eklund and Fred Isaksson, who together have had an equal part in the development of this report as well as building and testing the studied product. To work as efficiently as possible, responsibilities have been divided between the authors. When developing the chatbot, Fred was responsible for the conversation flow and content whilst Joakim was responsible for the backend functionality and integration. Although these responsibilities were set, a lot of interdisciplinary work was conducted by both authors. Also, when writing the report, specific chapters were divided between the authors for an effective writing process. Altogether the whole report was continuously revised and changed by both authors to maintain an equal quality all the way through.

Glossary

The following glossary defines central concepts in this thesis. Explanations of the concepts account for their respective meaning in the presented case and NOT as a general empirically based definition. Most concepts are elaborated and explained further later on.

Ability ~ A cognitive concept of trust in AS. Ability is a human's perception of the AS capability in performing expected tasks. This research treats the concept as the perceived reliability, capability and predictability of Ava.

AS ~ An acronym for "Autonomous System", and a reference to a system that is automated.

Automation ~ Enabling a process to run automatically without human intervention.

Ava ~ An acronym for "Artificial Vocational Advisor" and the provided name of the social chatbot service created and presented in this study.

Benevolence ~ A cognitive concept of trust in AS. Benevolence is a human's perception of the AS underlying positive intentions towards the human. This research treats the concept as the perceived prejudice, motives and beliefs of Ava.

Biases ~ The antonym to unbiasses.

Chatbot ~ A computer program which automatically conducts conversations via the means of different communication mediums.

Conversational Agent ~ A synonym for *Chatbot*. Occasionally shortened to *agent*.

HMI ~ An acronym for "Human-Machine-Interaction" and a reference to the interaction between a human and a machine.

Integrity ~ A cognitive concept of trust in AS. Integrity is a users perception of the AS loyalty to a set of principles that the human user has agreed upon. This research treats the concept as the perceived honesty, motives and character of Ava.

NLP ~ An acronym for "Natural Language Processing" and a reference to a script that enables the processing and interpretation of unsorted typed or spoken language, using machine learning.

Opaque/Opacity ~ The antonym to transparency. Opaque and Opacity are used in different grammatical situations.

Service Encounter ~ The moment in which a human for the first time interacts directly with the frontline of a service. In some disciplines referred to as the "moment of truth".

Social Chatbot ~ A computer program which automatically conducts socially oriented conversations via the means of different communication mediums.

System Component ~ A non-functional design aspect of the AS that is constituted by several functional design choices.

System Performance ~ In this research context, system performance is used as a general reference to a systems ability of performing expected tasks without faults or erroneous behaviour. Suggested to be one contributing system component, consisting of several individual design choices, to the perceived ability of Ava.

TIA ~ An acronym for "Trust In Automation" and a reference to the defined meaning of trust as a concept in the provided context of automation.

Transparency ~ In this research context, transparency is used as a general reference to the users overall possibility to view and understand the design, principles, functionality and limitations of a system. Suggested to be one contributing system component, consisting of several individual design choices, to the perceived integrity of Ava.

Trust ~ The overall psychological attitude achieved from beliefs and expectations about the AS trustworthiness. In this case, derived from the perceived integrity, benevolence and ability of a service encounter with a social chatbot, involving uncertainty and risk.

Unbiasses ~ In this research context, unbiasses is used as a general reference to objective and impartial system design choices. Optimising for mitigating the risk for including prejudisms and partisan beliefs in a certain context. Suggested to be one contributing system component, consisting of several individual design choices, to the perceived benevolence of Ava.