University of St Andrews

Department of Philosophy

Please do not modify the template styles
(Arial 12pt, left-align, 1.5 space)

ID NUMBER:	**230030434**
MODULE NAME:	**Reading Philosophy 1**
MODULE CODE:	**PY3100**
TUTOR'S NAME:	**Dr Miriam Bowen**
ESSAY TITLE:	**Essay**
WORD COUNT:	

I hereby declare that the attached piece of written work is my own work and that I have not reproduced, without acknowledgement, the work of another.

Proposal

(1) Working title: Trust in Artificial Agents

(2) Main topic / Main questions:

How does one define the trust a human has in an artificial agent?

How to define the trustworthiness of an artificial agent?

~~Is there a lack of rational basis in the trust that experts and the public commonly share?~~

(3) An indication of the main line you will take in the essay: what's the news going to be?

I shall argue that we cannot simply apply theories of trust between humans to the trust between humans and artificial agents. We need a separate set of theories for trust, or a generalised theory of trust that can describe both human agents and artificial agents.

~~I will address the question of whether there is a lack of rational basis in the trust that experts and the public commonly share. I will argue that while some aspects of trust in artificial agents are grounded in empirical evidence and expert analysis, other aspects may be influenced by irrational factors such as over-reliance on authority and fear of the unknown.~~

(4) Bibliography

The Routledge Handbook of Trust and Philosophy, Ch. 23-27

S. Kate Devitt, Trustworthiness of autonomous systems

~~Devesh Narayanan & Zhi Ming Tan, Attitudinal Tensions in the Joint Pursuit of Explainable and Trusted AI~~

Plan

Introduction

Technological optimist claims that the more comprehensively the verification is conducted, the more it can reduce the component of trust in the relationship between humans and AA since the reliability increases. I argue that the comprehensiveness of verification has only weak impact on the human trust. I shall further argue that how trustworthiness differs from mere reliability.

Previous scholarship argues that trust varies for different objects. There are attributes of trust that make trust relationships involving AAs distinct from those that only involve humans. I shall argue that we may not need to separate Human-AA trust and Human-Human trust; it is better to have a general trust theory.

Verification and Human-AA Trust

Verification: An answer to “does X conform to the specifications?”.
We can verify the correctness of AA under given circumstances. However, we often cannot guarantee that the AA will not encounter situations beyond its design, where behaviour becomes unpredictable. For example, Uber's test self-driving car fatally struck a joy walker because the engineers hadn't considered jaywalking possibilities. Verification is limited to the expectations of the designers, and comprehensive verification only confirms that the specifications come out of the expectations are fully satisfied. However, the object of trust is the AA, not the specifications.

Reliability and Trustworthiness of AA

A technological optimist might argue that verification establishes the reliability of AA.
However, performing consistently well is merely a quality of being trustworthy.
Methaphysical: What is trustworthiness of an AA?
Epistemological: How do we know an AA is trustworthy?

Trustworthiness is a quality of AA that potentially influences the Human-AA trust in a positive way. AA is to be made trustworthy to gain the human trust.

Human-Human Trust and Human-AA Trust Distinction

AA lacks human qualities and traits that are prerequisite for being an agent worthy of trust. This suggests that we cannot directly apply the theory of Human-Human Trust to Human-AA Trust; we need to establish a separate theory for Human-AA Trust.

Briefly explain the existing Human-Human Trust theories

1) Risk-assessment theories

2) Motives-based theories

3) Non-motives-based theories

Introduce Human-AA Trust theories

1) Taddeo’s e-Trust

3) Tavani’s model of trust

2) Grodzinsky, Miller, and Wolf’s Object-Oriented Model of Trust

Discuss these theories and reconsider the claim that trusting a human is different than trusting an AA.

Room for General Trust Theory

Three characteristics of trust: transparency, identity and predictability make trust relationships involving AAs distinct from those that only involve humans. Are they capable of making such a distinction?

The definition of trustworthiness may vary for different objects. Claim: Human trust may solely related to the trustworthiness of the object and is independent of what the object is, whether it's a Human, an AA, or an Alien.

Possible Responses

Conclusion