

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