



Calhoun: The NPS Institutional Archive
DSpace Repository

CRUSER (Consortium for Robotics and Unmanned Systems Education and Research)

2018-04-18

The Effect of Perceived Benevolence on Trust in Automation

Clark, Tiffany

<http://hdl.handle.net/10945/58056>

This publication is a work of the U.S. Government as defined in Title 17, United States Code, Section 101. Copyright protection is not available for this work in the United States.

Downloaded from NPS Archive: Calhoun



Calhoun is the Naval Postgraduate School's public access digital repository for research materials and institutional publications created by the NPS community. Calhoun is named for Professor of Mathematics Guy K. Calhoun, NPS's first appointed -- and published -- scholarly author.

Dudley Knox Library / Naval Postgraduate School
411 Dyer Road / 1 University Circle
Monterey, California USA 93943

<http://www.nps.edu/library>

The Effect of Perceived Benevolence on Trust in Automation

LT Tiffany Clark

Why Trust in Automation*?

- Human-machine teams are here/coming
- Not effective if human doesn't trust machine

* used broadly, up to & including fully autonomous machines with Artificial Intelligence (AI) software



Source: https://www.youtube.com/watch?v=_upbplsKGd4



Source: <http://author-quest.blogspot.com/2013/10/my-top-ten-favorite-robots.html>



Why Perceived Benevolence?

- Human-human trust typically broken by:
 - Competency-based errors (CBE)
 - “Oops! I forgot to include your work in my references”
 - Benevolent intent
 - Integrity-based violations (IBV)
 - “I purposely took your work and passed it off as my own”
 - Malevolent intent
- Regardless of actor’s actual intent, *perceived intent* breaks trust for victim

Assumption shift?

- General prior assumption:
 - automation does not/cannot have intent or integrity, so cannot commit IBV
- Reasons to challenge this:
 - Advances in AI (at human-level by 2050?)
 - AI learns from human behavior (including bad?)
 - AI programmer is human, so could easily have malevolent intent or lack of personal integrity
 - Cyber hacking skills can turn machines with benevolent designs into malevolent actors
- All leading to possible *perception* of intent

Thesis Experiment

- 106 participants in team-based exercise with AI teammate “BRIAN”
 - Task 1: Visual searches to earn points
 - Task 2: Choose to invest points in BRIAN, and BRIAN chooses how to share profits
 - CBE condition: error lost all invested points
 - IBV condition: BRIAN stole all invested points
- Measured trust response of participants by:
 - reliance behavior
 - investment amounts
 - response times
 - performance perceptions