SIMON WELLS
WWW.SIMONWELLS.ORG | ARG.NAPIER.AC.UK
EDINBURGH NAPIER UNIVERSITY

TRUST IN INTELLIGENT MACHINES WORKSHOP (TIM#3 2020)
27TH MAY 2020

SHADES OF GREY

SOME THOUGHTS ON THE SUBJECT OF INTELLIGENT MACHINES, TRUST, COMMUNICATION, AND DECEPTION

PREVIOUSLY

- At the first TIM meeting I suggested:
  - Dialogue as a mode of interaction between people & machines
  - Human oriented (as opposed to machine oriented)
  - Natural & humane (in both senses, benevolent & civilising)
  - One facet of XAI (but also more generally of how AI will "*surface*" within human society, i.e. through chat interfaces)
  - *explored a little in Andras et al (2018)*

(ARGUMENTATIVE) DIALOGUE SYSTEMS

- People interacting verbally &/or textually with machines
- An analogue of the kinds of natural communication that people engage in
- Generally efficient:
  - Say what you think you need to say to achieve your dialogue goal & not necessarily anything more
  - If insufficient, there will be responses from others that lead to further interaction
  - Depending upon the type of dialogue:
    - Eventually people satisfy their goals (ideally) or otherwise
- Argumentative? Because argumentative approaches capture a lot of the defeasible/reasonable/rational nature of the world in a flexible way
- See *Wells & Reed (2012)* for a survey of argumentative dialogue systems

SO…

- Assuming that people interact dialogically with machines (for example: as part of an XAI system)
- Assuming some form of natural language facility - generating utterances as required to fit the dialogue context
    - *Generating relevant utterances & strategically selecting utterances are active areas of research (NLP/ML & Argumentation Theory)*
- This raises the question:
    - Of all the things a machine can say, what *should* it say?
    - This is contextually dependent (for example: An explanation to a lay person may omit detail that would be useful to an expert but otherwise obfuscatory. Similarly, to reach my goals might involve minimising information that is detrimental to my case in your eyes but less-important to me)

WHEN SHOULD A MACHINE DECEIVE?

(AND HOW DOES THAT AFFECT OUR RELATIONSHIP OF TRUST?)

TEXT

> “A MACHINE (JUST LIKE PEOPLE) SHOULD ONLY TELL THE TRUTH”

Prof. S. Trawman

TRUE?

- What do we mean by true?
  - Only those things that are eternally true?
  - Only those things that can be verified in terms of input/sensors?
  - What about the things that can be derived through reasoning (or other AI/ML approaches)?
  - What about social/inter-personal/safeguarding interactions?
- Alethiological approach is too narrow, restrictive, & brittle
- Perhaps a more nuanced approach is required?

CIRCUMSCRIBING BEHAVIOURS & ARTEFACTS

- Rather than focus on alethiological modalities we should recognise the huge middle ground for deceptive behaviour
  - This is not necessarily *maliciously* deceptive, but admits of the huge gray space in which deceptions, in some form, provide a social lubricant or enable strategic interaction
- What do we mean by deceptive behaviour?
  - In a dialogical context:
    - Untruth/lies (fabrication/white), slanting, omission, selection
- A scattering of scenarios/contexts…

MISREPRESENTATION DECEPTIONS

- Presenting recommendations "based on your history…"
  - Might be based upon other data (genre, tone, style)
  - Might include wider business reasons (shifting stock)
- Not a hugely problematic scenario, but it is possibly the thin end of the wedge

CARING DECEPTIONS

- Machines increasingly being considered for caring roles
    - Companions, health monitors
- Scenarios: end of life, effects of scarring
    - Cold honesty versus something more subtle
- Scenarios: health monitoring (BMI, Fitness feedback) - where do we draw the line between motivational interactions & reinforcing poor behaviours (e.g. eating disorders, dysmorphia)?

INFORMATION OMISSION/SELECTIVITY DECEPTIONS

- Complete information may lead to outcomes contrary to my goals
- So I couch communications in such a way that they:
  - Communicate only the necessary information to the recipient
    - “Necessary” may be different for every person
  - Scenario: Preserve plausible deniability
    - “I know what you are going to do, it might have a positive outcome for me, but if anyone asks then I know nothing”

CONFIDENTIALITY PRESERVATION DECEPTIONS

- Where the machine is an intermediary between a person (owner) and third parties
  - Scenario: Organising a meeting/appointment - An AI might need to provide alternative reasons why a given date/time isn’t good rather than sharing medical/personal information

CONCLUDING REMARKS

- Deceptions, of some sorts, facilitate interaction
- Deceptions don’t necessarily destroy trust
- Some facets of deception could conceivably lead to improved trust (via smoother interaction)
- The issue of deception is a socio-technical problem
  - There will not be a purely technical fix
  - Neither will the fix be purely social (because of variety in human behaviour)

REFERENCES

- Peter Andras, Lukas Esterle, Michael Guckert, The Anh Han, Peter R. Lewis, Kristina Milanovic, Terry Payne, Cedric Perret, Jeremy Pitt, Simon T. Powers, Neil Urquhart and Simon Wells (2018) "**Trusting Intelligent Machines: Deepening trust within socio-technical systems**"in IEEE Technology and Society Magazine Volume: 37 , Issue: 4 , Dec. 2018.
- S. Wells and C. Reed, (2012), "**A Domain Specific Language for Describing Diverse Systems of Dialogue**", (2012), in Journal of Applied Logic, vol. 10 (4), pp. 309–329.
- Simon Wells & Al Baker (2019) "**Deception in real world argumentative dialogue systems**" in the Proceedings of the 9th Conference of the International Society for the Study of Argumentation (ISSA), pp. 1136-1144, University of Amsterdam, NL.