MIND THE GAPS

Deception in real world argumentative dialogue systems

Coming Up:

- AI & People
- Dialogue: Humane Interfaces, Real-world Human-Machine Dialogue, XAI
- Context of Dialogue
- Problematic Cases
- Some aspects of an ethical framework

Capabilities, Delegation, & Trust

- Trends
  - Increases in capabilities of intelligent machines
  - Increasing delegation of decision making to machines
- Issues
  - Societal need for explanation (parity of treatment)
  - Legal mandates for scrutinisation & interpretation (regulated sectors)
  - Socio-technical complexity (room for conflict)
  - Trust (“computers says no” versus “This black box denied my loan”)

Dialogue: A Humane Interface

- Natural interface
- We understand & trust by exploring and explaining
- We build confidence by justifying
- Acceptability of an explanation can be dependent upon target (can easily shift into justification)
- Tendency for humans to mistrust anything different
- Need for trust if machines are to act effectively within society

Real-world Dialogue Application

- Modelling Dialogues as normative formal systems [**Hamblin (1970), Walton & krabbe (1995)**, + many more]
- Computational Approaches:
  - MAgtALO - MultiAgent Argument Logic & Opinion [**Reed & Wells (2006)**]
  - DGDL - Dialogue Game Description Language [**Wells & Reed (2012)**]
  - ADAMANT - A DiAlogue MANagement Tool [**Wells (*forthcoming*)**]
- A dialogical interaction system can support both explanatory & justificatory modes of communication between people & machines in a humane fashion

Explainable AI (XAI)

- Developers of ML systems are reinventing notions of explanation for their own purposes (**Doran (2018)**)
- How to achieve explanations - train a second AI to explain the first (amongst other approaches: **Gregor & Benbasat | Gunning | Ribeiro | Oren *et al***)
- Explanation means different things to different groups - dialogues are contextual
  - System Designer/Engineer
  - System Owner
  - System Controller
  - End User (direct)
  - End User (indirect)
  - Legislative, Licensing, Legal
  - <u>Some of these contexts could give rise to deceptive interactions</u>
- Interesting and fundamental research on AI but different communities unaware of each other

Contexts of Dialogue

- Assuming we can build AI systems that can explain their reasoning in various contexts
- Opportunities for strategic reasoning - (Searle “where there’s choice there’s strategy”)
- Opportunities for deception
    - Typology of lies [Kraus (yesterday)]
    - Huge body of work on deceptive & related practises (from lies of omission, through framing, bias)
    - With heterogeneous ownership of & interaction between machines & people -
- “When should a machine lie or otherwise deceive?”
- How to handle this? Mechanisms, Education, Law, Ethics, …?

Some Deceptive Cases

- 1) Pro-Social explanations of AI behaviour
  - (always?) Good. At least useful
- 2) Deceptive explanations to regulatory bodies
  - (always bad)
- 3) Recommender system explanations
  - (edge cases, hmmmm.)

Key questions

- How do we account for the intuitive moral difference between these cases?
- Is that difference applicable in robot to human deception?

- Can Robots Deceive and/or lie?
- If robot lies are wrong, are they wrong for the same reasons as human lies?

Ethics of Lying and Deception

- If some lies are morally permissible, how do we tell the difference?
- In other words, when a lie is wrong, what's wrong with it?

"*In our own time we find it particularly natural to think deceiving people (or at least some people, in some circumstances) is an example of using or manipulating them, and that that is what is wrong with it.*" **Bernard Williams**

So: **Lies** can be wrong because and just in case they are **deceptive**, and deception can be wrong because and just in case they are **manipulative**.

Lying and Deception

- Lying is one of many forms of deceptive communication.
- To deceive $=_{df}$ to intentionally cause to have a false belief that is known or believed to be false.
- The definition of lying continues to be contested, but a traditional definition might be the following:

> A lie is a **statement**, which the speaker **believes to be false** made with the **intention** that a **person** should **believe the statement to be true.** (Mahon 2015)

- On this definition:
  - Lies are not necessarily deceptive
  - Lies are not necessarily morally impermissible
  - Deception by omission is not a lie

Other forms of deception

- ***Paltering***

  Deceiving by uttering an (irrelevant) truth: e.g. Saint Athanasius, who replying to his persecutors' question 'Where is the traitor Athanasius?' with the misleading truth 'Not far away.'

- ***Omission***

  Causing someone to form a false belief by withholding relevant information that would challenge it.

- ***Bullshit***

  An attempt to persuade without regard for truth.

- ***Double Bluff***

  Making a truthful utterance with the intention that an interlocuter believes you to be deceptive.

Can Robots Deceive?

- Can they lie?
  - Can they **make statements**
  - Can they **believe things to be false**
  - Can they **intend an interlocutor to form a belief**

- **In Principle** I think we can grant these, or close analogues.
- **In Practice** a question hangs over the intention condition: note that this is essential to *all* forms of deception.

Why Are Lies Manipulative?

Because they breach ***Trust:*** the kind of trust that lends testimony an assurance, beyond merely being evidence for the speakers’ belief, that the belief is warranted in the listener.

- Predictive Trust

  (1) A knowingly depends on S ϕ-ing and

  (2) A expects S to ϕ (where A expects this in the sense that A predicts that S will ϕ).

- Affective Trust

  (1) A knowingly depends on S ϕ-ing and

  (2) A expects S’s knowing that he depends on S ϕ-ing to motivate S to ϕ

(Faulkner 2007)

Consequences

- If Robots cannot value humans in the right kind of way, is it reasonable for us to expect that our dependence on them is what motivates their actions?

- If Robots do *not* have that capacity, we would be making a mistake if we were to blame them for betrayal (as when I hit my printer for betraying me by making me late).

- We cannot blame the robots, we cannot easily blame the designers/engineers, so the wrongness of AI deception cannot be located in the same place as human deception.

- It looks like we plausibly have a *Responsibiliy Gap,* but of a distinctive and unsettling kind.

To conclude…

- Mechanical solutions are challenging
- Educational solutions are long term (& challenging)
- Law is progressing rapidly
- Preliminaries of an ethical framework (but challenging)

- Things are progressing on many fronts
- There are still theoretical & applied gaps