WASP-HS AI for Humanity and Society 2024
Workshop

IS AI DECEPTION DECEPTION?

PROGRAM:

INTRODUCTION & WELCOME (0855 Hrs. – 0900 Hrs.)	
SESSION 1 – WHAT IS AI DECEPTION?	
INVITED TALK Dimitri Coelho Mollo (Umeå University) *Human-sounding chatbots: dignity and deception*	0900 Hrs. – 0935 Hrs.
LIGHTNING TALK Kesavan Thanagopal (University of Notre Dame) *Deceptive AI or Deceptive Framing?*	0935 Hrs. – 0945 Hrs.
INVITED TALK Jessica Pepp (Uppsala University) *Manipulative machines*	0945 Hrs. – 1020 Hrs.
BREAK (1020 HRS. – 1035 HRS.)	
SESSION 2 – WHAT TO DO ABOUT AI DECEPTION?	
INVITED TALK Ondřej Krása (University of Pardubice) *Preventing Deceptive Alignment?*	1035 Hrs.– 1110 Hrs.
LIGHTNING TALK Maria Zanzotto (Università degli Studi di Torino) *Manipulation from non-generative AI to generative AI: the issue of indistinguishability*	1110 Hrs.– 1120 Hrs.
INVITED TALK Rachel Sterken (University of Hong Kong) *Evaluating and Mitigating AI Manipulation and Deception*	1120 Hrs.– 1155 Hrs.
CONCLUDING REMARKS (1155 HRS. – 1200 HRS.)	

SESSION 1 – WHAT IS AI DECEPTION?

INVITED TALK: DIMITRI COELHO MOLLO (UMEÅ UNIVERSITY)

Title: Human-sounding chatbots: dignity and deception

> Thanks in part to finetuning techniques such as Reinforcement Learning through Human Feedback (RLHF), current chatbots produce outputs that display many of the linguistic cues associated with agency, rational and emotional attitudes, and even consciousness. However, there is little doubt that chatbots today lack these features.
>
> In this talk, I will present co-authored work investigating the ethical risks that human-sounding chatbots pose to users. I will focus especially on risks connected to deception, misplaced trust, and potential offences to human dignity.

Suggested Reading: https://arxiv.org/abs/2406.18346

Some Further Research Questions:

1. Can there be deception without intention to deceive? E.g., do chatbot designers need to intend to deceive in order to create a deceitful chatbot?
2. What are the boundaries between make-believe and deception in interactions with chatbots?
3. How should we look for the balance, if it exists, in the trade-off between user-friendliness and potential deception?

LIGHTNING TALK: KESAVAN THANAGOPAL (UNIVERSITY OF NOTRE DAME)

Title: Deceptive AI or Deceptive Framing? A Peek Behind the Proverbial Wizard's Digital Curtain

> Current conversations about AI deceptions might themselves be deceptive in at least two different ways. First, we often fail to distinguish between (i) cases of agential deception using AI systems (henceforth referred to as "agential deception") where AI systems are used, either wittingly or unwittingly, by agents – nefarious or otherwise – as tools for deception, and (ii) cases where the AI systems appear to be the principal source of deception without the direct interference of any particular agent (henceforth referred to as "artificial deception"). The conflation of these two types of AI deceptions obscures the distinctive nature of each type of deception, tempting us to think, rather mistakenly, that a "unified resolution" in the form of some kind of regulatory framework exists to mitigate "AI deception" as a whole. Second, those who do distinguish agential deception from artificial deception nevertheless often appeal to a kind of agency – an "artificial agency" – that the corresponding AI system possesses in discussing artificial deception, anthropomorphising AI systems as entities that engage in "deceptive behaviour".
>
> The goal of this lightning talk is rather modest. I will begin by first elucidating the difference between agential deception and artificial deception. In so doing, I aim to clarify why a unified resolution of the regulatory kind to "AI deception" would ultimately be found to be lacking. I will then hone in on the notion of artificial deception and explain why anthropomorphising AI

systems as artificial agents might make us more susceptible to forming false beliefs through our interactions with them, hindering any attempts to mitigate this particular type of deception. If I am right, then what lies behind the proverbial wizard's "digital curtain" isn't so much an artificial agent trying to outsmart and deceive us at all; rather it is an illusionary authoritative figure we have unwittingly self-constructed, leading us to deceive ourselves when we interact with such AI systems.

INVITED TALK: JESSICA PEPP (UPPSALA UNIVERSITY)

This talk will address the following questions:

What is the difference between AI deception and AI manipulation? Do we need the concept of manipulation in order to articulate and address important concerns about the behaviour of AI systems? If we do, which conceptions of manipulation might be used to do this? Do ordinary conceptions of manipulative behaviour apply to AI systems? For there to be manipulation, must there be manipulative intent?

Suggested Reading:

https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781003205425-6/manipulative-machines-jessica-pepp-rachel-sterken-matthew-mckeever-eliot-michaelson

Some Further Research Questions:

1. To what extent is non-manipulation a goal for a value-centred approach to design for AI systems?
2. In aiming to produce "non-manipulative" AI, to what extent should we be guided by "folk concepts" of manipulation, as opposed to revised notions purpose-built to achieve certain other purposes?
3. How can manipulation be identified in today's and tomorrow's AI systems?

SESSION 2 – WHAT TO DO ABOUT AI DECEPTION?

INVITED TALK: ONDŘEJ KRÁSA (UNIVERSITY OF PARDUBICE)

Title: Preventing Deceptive Alignment?

Deceptive alignment represents a critical challenge in AI safety (Cotra, Hobbhahn): If powerful AI systems manage to deceive examiners about their safety, they could pose significant risks to humanity once deployed. There is already evidence that current AI models can exhibit deceptive alignment in testing environments (OpenAI).

Researchers have proposed several strategies to prevent models from deceiving us. One promising and popular approach is to identify the parts of the neural network responsible for deceptive behaviours (e.g., features related to honesty) and manipulate them to encourage honesty in the model (Templeton, Zou). However, this approach faces significant conceptual obstacles.

The first challenge lies in the complexity of deception itself. Deception may be a multifaceted concept, with different forms potentially corresponding to different parts of the neural network. Identifying a feature associated with a specific instance of deception does not necessarily ensure that it applies to other types of deceptive behaviour. This complexity is supported by both neuroscientific studies of human deception (Ganis) and recent analyses of the truthfulness of large language models (Orgad).

The second challenge builds on this complexity. The foundation of efforts to understand key concepts within artificial neural networks is the assumption that AI models learn human-comprehensible algorithms (Nanda). However, if we consider that AI systems may surpass human intelligence in domains related to deception (Meta, Hagendorff), it might become fundamentally impossible for us to fully understand some critical aspects of deception. This would mean that identifying certain forms of AI deception may be beyond our grasp, as it is self-contradictory to expect humans to comprehend reasoning that exceeds human intelligence. In other words, we may be "simply not smart enough to understand certain concepts." (Yampolskiy)

LIGHTNING TALK: MARIA ZANZOTTO (UNIVERSITÀ DEGLI STUDI DI TORINO - FINO CONSORTIUM)

Title: *Manipulation from non-generative AI to generative AI: the issue of indistinguishability*

In recent years, a compelling narrative has emerged, positioning Generative AI as a transformative force reshaping society, business, and daily life. However, less effort devoted to examining *how* such change might unfold. My research aims to explore how generative AI technologies may fundamentally alter existing issues and introduce new ethical concerns when compared to non-generative models like traditional machine learning.

A core focus of my work is on manipulation. In non-generative models, one notable instance of manipulation was the Cambridge Analytica scandal, where Facebook user data powered algorithms that were meant to targeted and influenced swing voters. In contrast, a defining

feature of generative AI is its ability to produce content that is increasingly indistinguishable from human-made outputs. This indistinguishability has profound implications for manipulation, especially through AI-generated images, videos, texts, and chatbot interactions that deeply reshape online engagement.

But what makes "indistinguishability" manipulative? This talk, building on Ienca's (2023) notion of manipulation, examines whether interaction with AI-generated content can diminish personal autonomy, referencing Coeckelbergh's (2023) insights on epistemic agency. Central to this discussion is the relational dimension of human-AI interaction. Drawing on Dennett's intentional stance, we tend to attribute beliefs, intentions, and emotions to conversational partners, yet with AI "agents," these perceived qualities exist only in the observer's mind. Despite our assumptions of meaning and intention, we are interacting with what Bender, Gebru, McMillan-Major and Shmitchell have called "stochastic parrots" (2021). Dennett himself expressed concern over interactions with synthetic agents in "The Problem with Counterfeit People" (2023).

This talk opens the discussion to whether these interactions are inherently manipulative and if they introduce new vulnerabilities, or if, instead, they prompt us to reconsider and redefine relational boundaries in our digital age.

INVITED TALK: RACHEL STERKEN (UNIVERSITY OF HONG KONG)

Title: Evaluating and Mitigating AI Manipulation and Deception

As AI systems become more pervasive and powerful, the potential for AI-driven deception and manipulation presents significant practical challenges. This presentation explores methods for evaluating AI systems' capabilities for deception and manipulation, and strategies for mitigating AI systems' deceptive and manipulative behaviors. Given the early stage of research in this area, the talk will focus on the inadequacy of current methods and strategies, and point to open questions in the area.