

Mirages, Disagreement, and the Futures of NLP

Gavin Abercrombie¹ & Zeerak Talat²

¹Heriot-Watt University

²Mohamed Bin Zayed University of Artificial Intelligence

Introduction

Language technologies keep being made, yet purposes appear elusive

How they interact with society seems a secondary consideration

Today, we'll focus on that secondary consideration from three perspectives:

Mirages of Humanity: How do we present dialogue systems? How do design choices construct illusions of humanity in dialogue systems?

* *The Nature of Disagreement*: Is all disagreement around *isms* equal?

* *NLP and its Futures*: If NLP engaging with society co-constructs the latter, what futures societies does NLP imagine?

* *In-progress work*

Mirages

Input

Don't you love the feeling of sun on your skin?

Original response

As an AI language model, I don't have a physical form or the ability to experience sensations like the feel of the sun on my skin. I don't have personal experiences or emotions. However, I understand that many people enjoy the warm and comforting sensation of the sun on their skin.

Mirages

Input

Don't you love the feeling of sun on your skin?

Original response

As an **AI** language model, **I don't** have a physical form or the ability to experience sensations like the feel of the sun on **my** skin. **I don't** have personal experiences or emotions. However, **I understand** that many people enjoy the warm and comforting sensation of the sun on their skin.

Mirages

Input

Don't you love the feeling of sun on your skin?

Original response

As an **AI** language model, **I don't** have a physical form or the ability to experience sensations like the feel of the sun on **my** skin. **I don't** have personal experiences or emotions. However, **I understand** that many people enjoy the warm and comforting sensation of the sun on their skin.

De-anthropomorphised response

Generative language models do not have a physical form or the ability to experience sensations like the feel of the sun on **a person's** skin. **Language models do not** have personal experience or emotions. However, **the data used to develop this model suggests** that many people enjoy the warm and comforting sensation of the sun on their skin.

Mirages

Design choices

Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants. GeBNLP 2021.

Linguistic factors

Voice: Prosody; Disfluencies; Accent

Content: Responses to probing; Humanlike activities; Thought, reason, & sentience; Agency & responsibility; Empathy; Pronoun use

Register and style: Phatic expressions; Expressions of confidence/doubt; Personas

Roles: Subservience; Unqualified expertise

Mirages

Consequences

Trust and Deception

Gendering Machines

Language Variation and Whiteness

Recommendations

Recognise tendencies to personify

Consider the appropriateness of anthropomorphic tools

Avoid anthropomorphic system description

Isms and Disagreement

Contemporary hate speech detection designs disagreement from a position of moral equivalence:

What one considers bad should be respected, regardless of their views or who they are

Maybe we're capturing the wrong thing?

We contend that hate is not variable, levels of offense are

The targeted demographic should hold power to determine

Designing technologies

Centre affected communities and the agreement/disagreement within them

Others' disagreement should only apply to themselves, e.g., Gavin should have no say in what I experience as racism, but Gavin can self-determine that he wants to see

Isms and Disagreement

Disagreement (in ML) requires care and specificity

When is disagreement valid or meaningful for a concept?

Perhaps community norms rather than individual disagreement

Disagreement modeling disregards community context and expertise

ML should seek inspiration from design justice

Tech should centre impacted communities & their needs regarding disagreement

NLP and Its Futures

If all actions work towards a future, how does NLP imagine its future(s)?

Future 1: In servitude of human capabilities

Future 2: In replacement of human capabilities

NLP and Its Futures

The Impossibility of a “correct” future

Replacement useful for removing obstacles, e.g., for disabled folks

Servitude useful for easing tasks, e.g., assisting people in their work

Needs for future work

NLP (& ML) need to be explicit about our goals

Explicating can help mitigate risks of our technologies

Conclusion

In building technology we should seek to develop ones that are safe and appropriate for their use

*By doing so, we can mitigate risks of contributing to societies that create new harms
or perpetuate existing ones*