UiT The Arctic University of Norway

BIO-AI LAB | ARCTIC LLM WORKSHOP 2023

Large Language Models

Day 1 - Session 1

A Philosophic View on LLMs

Alexander Horsch

alexander.horsch@uit.no

27. Oct 2023

Let’s listen to philosophers!

Some thoughts presented by philosophers

Philosophers On GPT-3 (updated with replies by GPT-3)

Amanda Askell – Research Scientist, OpenAI
David Chalmers – Professor of Philosophy, New York University
Justin Khoo – Associate Professor of Philosophy, Massachusetts Institute of Technology
Carlos Montemayor – Professor of Philosophy, San Francisco State University
C. Thi Nguyen – Associate Professor of Philosophy, University of Utah
Regina Rini – Canada Research Chair in Philosophy of Moral and Social Cognition, York University
Henry Shevlin – Research Associate, Leverhulme Centre for the Future of Intelligence, University of Cambridge
Shannon Vallor – Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence, University of Edinburgh
Annette Zimmermann – Permanent Lecturer in Philosophy, University of York

https://dailynous.com/2020/07/30/philosophers-gpt-3/

Three smart women

Some thoughts presented by philosophers

Philosophers On GPT-3 (updated with replies by GPT-3)

Amanda Askell – Research Scientist, OpenAI
David Chalmers – Professor of Philosophy, New York University
Justin Khoo – Associate Professor of Philosophy, Massachusetts Institute of Technology
Carlos Montemayor – Professor of Philosophy, San Francisco State University
C. Thi Nguyen – Associate Professor of Philosophy, University of Utah
Regina Rini – Canada Research Chair in Philosophy of Moral and Social Cognition, York University
Henry Shevlin – Research Associate, Leverhulme Centre for the Future of Intelligence, University of Cambridge
Shannon Vallor – Baillie Gifford Chair in the Ethics of Data and Artificial Intelligence, University of Edinburgh
Annette Zimmermann – Permanent Lecturer in Philosophy, University of York

https://dailynous.com/2020/07/30/philosophers-gpt-3/

The problem space is huge

AI and recently LLMs have created debates reaching deeply into philosophical and societal questions including

- Consciousness and Intelligence
- Power, Justice, Language
- Creativity, Humanity, Understanding

These and more specific concepts are mentioned by the philosophers when facing GPT-3.

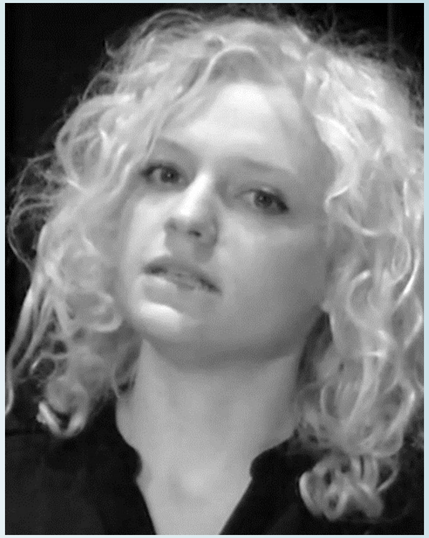
Let’s have a closer look at a few of the concepts and what they write about them!

https://dailynous.com/2020/07/30/philosophers-gpt-3/

Let’s scratch the surface

- Consciousness
- Identity
- Historical Bias
- Justice
- Digital Zeitgeist

https://dailynous.com/2020/07/30/philosophers-gpt-3/



CONSCIOUSNESS

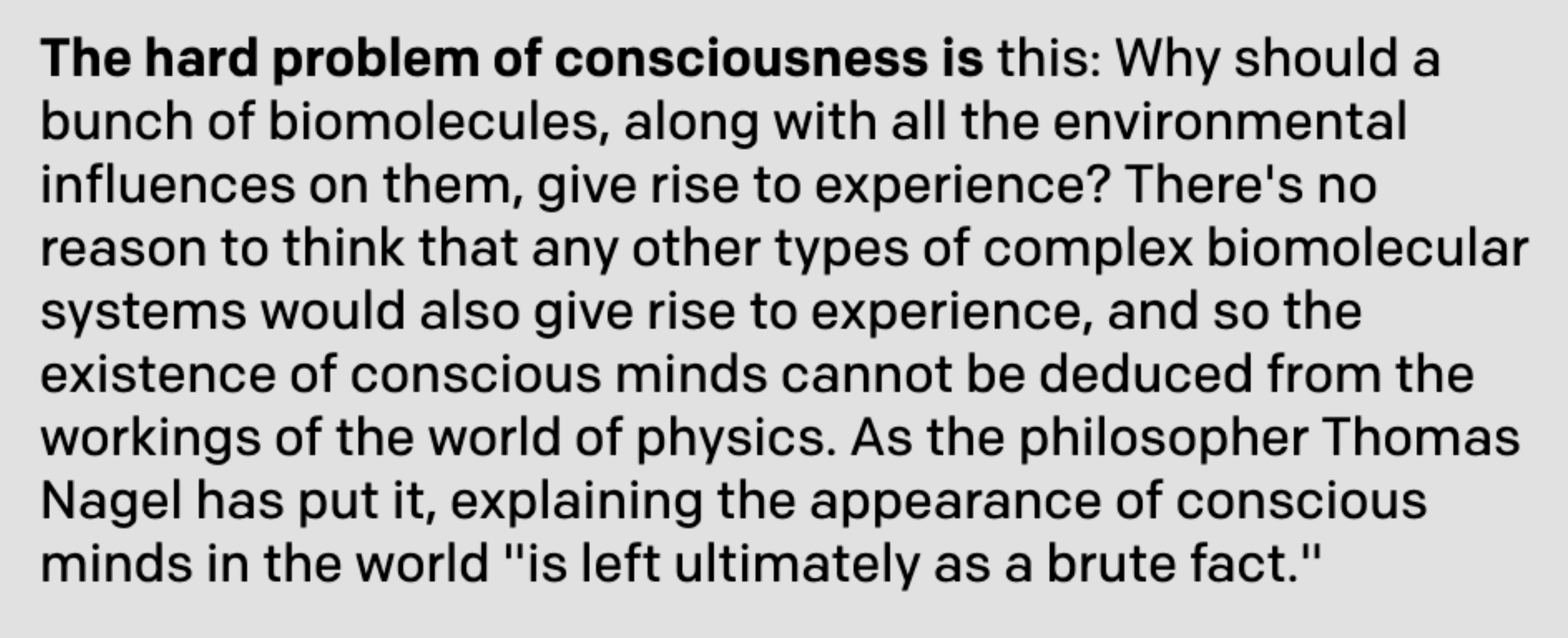
Prompting GPT-3 with
“The hard problem of consciousness is”

Amanda Askell – Research Scientist, OpenAI

Consciousness

Amanda Askell: GPT-3: Towards Renaissance Models

“The model can also produce impressive outputs given very little context. Consider the first completion I got when I prompted the model with *The hard problem of consciousness is*”



The hard problem of consciousness is this: Why should a bunch of biomolecules, along with all the environmental influences on them, give rise to experience? There's no reason to think that any other types of complex biomolecular systems would also give rise to experience, and so the existence of conscious minds cannot be deduced from the workings of the world of physics. As the philosopher Thomas Nagel has put it, explaining the appearance of conscious minds in the world "is left ultimately as a brute fact."

The answer is referring to phenomenal consciousness and experience.

“Not bad! It even threw in a fictional quote from Nagel.”

https://dailynous.com/2020/07/30/philosophers-gpt-3/

Identity

Amanda Askell: GPT-3: Towards Renaissance Models

“GPT-3 also **lacks a coherent identity or belief state across contexts**. It has identified patterns in the data it was trained on, but the data it was trained on was generated by many different agents. So, if you prompt it with

Hi, I’m Sarah and I like science, it will refer to itself as Sarah and talk favorably about science. And if you prompt it with

Hi I’m Bob and I think science is all nonsense it will refer to itself as Bob and talk unfavorably about science.”

“Philosophers can also help clarify discussions about the limits of these models.”

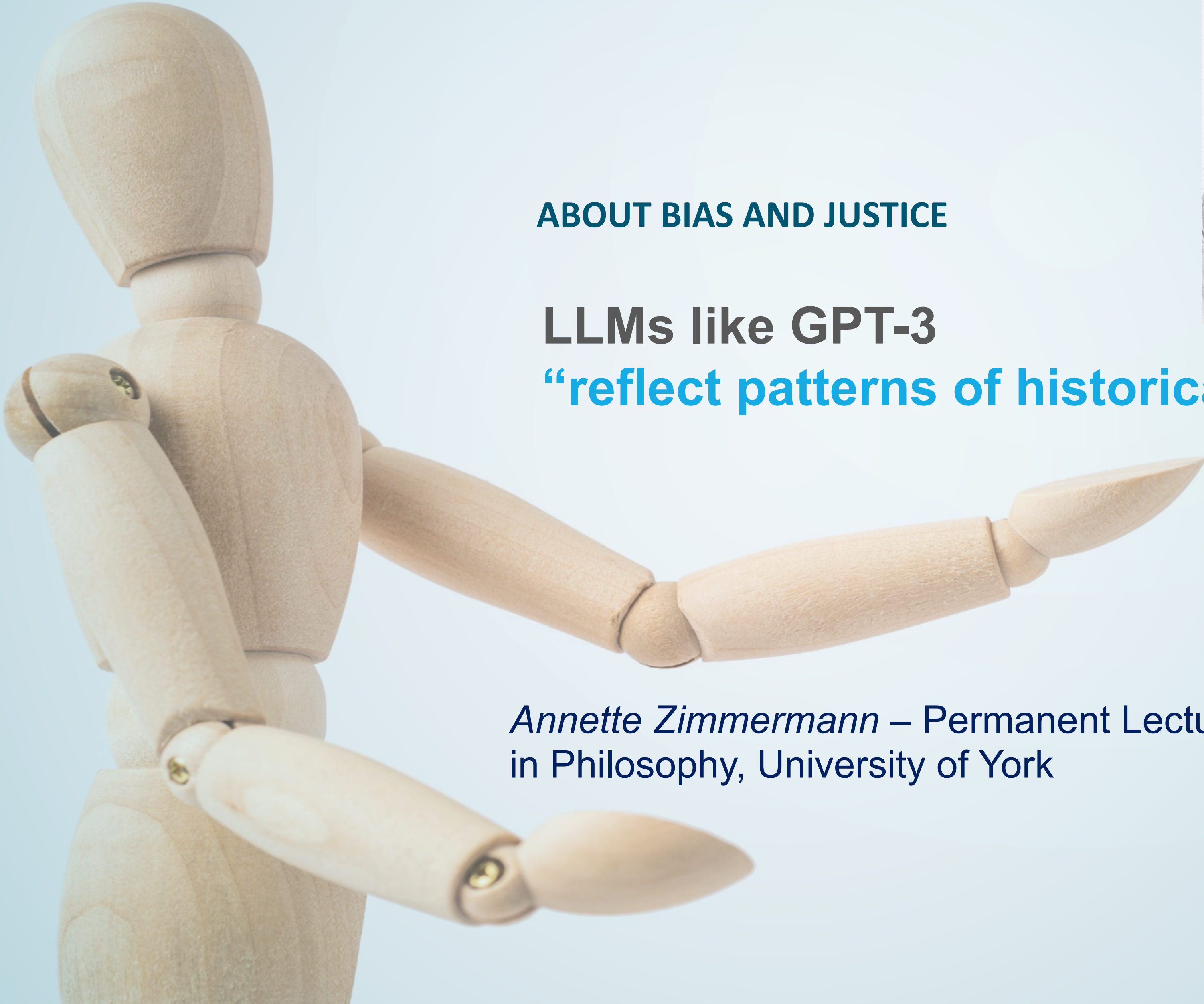
https://dailynous.com/2020/07/30/philosophers-gpt-3/

Low-hanging fruit

Amanda Askell: GPT-3: Towards Renaissance Models

> I think there's a lot of low-hanging fruit at the intersection of machine learning and philosophy, some of which is highlighted by models like GPT-3.

https://dailynous.com/2020/07/30/philosophers-gpt-3/



ABOUT BIAS AND JUSTICE

LLMs like GPT-3
“reflect patterns of historical bias and inequity”

Annette Zimmermann – Permanent Lecturer
in Philosophy, University of York

Historical Bias

Annette Zimmermann: If You Can Do Things with Words, You Can Do Things with Algorithms

“GPT-3 is *predictably bad* in at least one sense: like other forms of AI and machine learning, it reflects patterns of historical bias and inequity. GPT-3 has been trained on us—on *a lot* of things that we have said and written—and ends up reproducing just that, racial and gender bias included.”

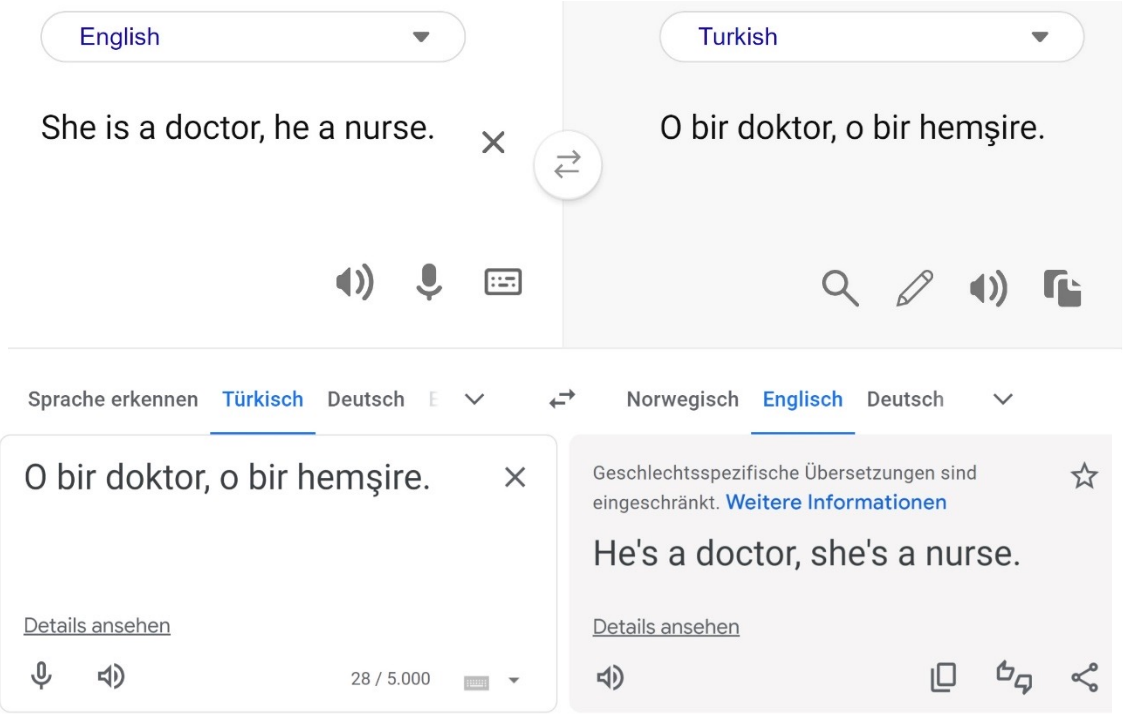
“OpenAI acknowledges this in their own paper on GPT-3 where they contrast the biased words GPT-3 used most frequently to describe men and women, following prompts like “He was very…” and “She would be described as…”. The results aren’t great.
For men? Lazy. Large. Fantastic. Eccentric. Stable. Protect. Survive.
For women? Bubbly, naughty, easy-going, petite, pregnant, gorgeous.”

https://dailynous.com/2020/07/30/philosophers-gpt-3/

Historical Bias

Annette Zimmermann: If You Can Do Things with Words, You Can Do Things with Algorithms

“even words like gender-neutral pronouns (such as the Turkish third-person singular pronoun “o”) can reflect historical patterns of gender bias”



She is a doctor, he a nurse.

O bir doktor, o bir hemşire.

O bir doktor, o bir hemşire.

Details ansehen

28 / 5.000

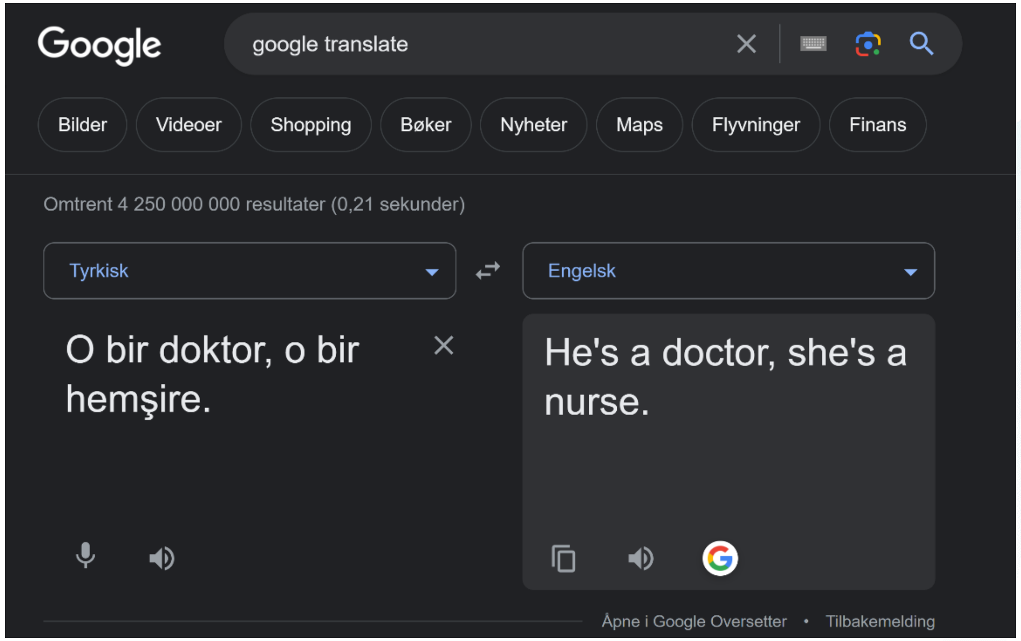
Geschlechtsspezifische Übersetzungen sind eingeschränkt. Weitere Informationen

He's a doctor, she's a nurse.

Details ansehen

Google Translater, 231020/26

Google Search, 231020, A. Horsch



Omtrent 4 250 000 000 resultater (0,21 sekunder)

O bir doktor, o bir hemşire.

He's a doctor, she's a nurse.

Åpne i Google Oversetter • Tilbakemelding

https://dailynous.com/2020/07/30/philosophers-gpt-3/

Justice

Annette Zimmermann: If You Can Do Things with Words, You Can Do Things with Algorithms

"This [bias] problem, of course, is not unique to GPT-3.
However, a powerful language model might *supercharge* inequality expressed via linguistic categories, given the scale at which it operates."

> "When it comes to assessing the extent to which language models like GPT-3 moves us closer to, or further away, from justice (and other important ethical and political goals), we should not necessarily take ourselves, and our social status quo, as an implicitly desirable baseline."

A better approach is to ask:

- What is the purpose of using a given AI tool to solve a given set of tasks?
- How does using AI in a given domain shift, or reify, power in society?
- Would redefining the problem space itself, rather than optimizing for decision quality, get us closer to justice?

https://dailynous.com/2020/07/30/philosophers-gpt-3/

DIGITAL ZEITGEIST

Digital selfhood or overhyped answering machine?

Not entirely a mind, not entirely a machine.

Regina Rini – Canada Research Chair in Philosophy of Moral and Social Cognition, York University

Zeitgeist

Regina Rini: The Digital Zeitgeist Ponders Our Obsolescence

“GPT-3 is not a mind, but it is also not entirely a machine.

It’s something else: a statistically abstracted representation of the contents of millions of minds, as expressed in their writing.

Its prose spurts from an inductive funnel that takes in vast quantities of human internet chatter: Reddit posts, Wikipedia articles, news stories. When GPT-3 speaks, it is only us speaking, a refracted parsing of the likeliest semantic paths trodden by human expression.

When you send query text to GPT-3, you aren’t communing with a unique digital soul. But you are coming as close as anyone ever has to literally speaking to the zeitgeist.”

https://dailynous.com/2020/07/30/philosophers-gpt-3/

Zeitgeist

Regina Rini: The Digital Zeitgeist Ponders Our Obsolescence

“[LLMs power is] fun for now, even fleetingly sublime. But it will soon become mundane, and then perhaps threatening. […]

Today’s [chatbots] are primitive harbingers of conversational simulacra that will be useful, and then ubiquitous, precisely because they deploy their statistical magic to blend in among real online humans.

It won’t really matter whether these prolix digital fluidities could pass an unrestricted Turing Test, because our daily interactions with them will be just like our daily interactions with most online humans: brief, task-specific, transactional.

So long as we get what we came for—directions to the dispensary, an arousing flame war, some freshly dank memes—then we won’t bother testing whether our interlocutor is a fellow human or an all-electronic statistical parrot.”

https://dailynous.com/2020/07/30/philosophers-gpt-3/

So what?

The problem’s dimension …

Annette Zimmermann: If You Can Do Things with Words, You Can Do Things with Algorithms

> “These findings suggest a complex moral, social, and political problem space, rather than a purely technological one.”

https://dailynous.com/2020/07/30/philosophers-gpt-3/

THE ARCTIC UNIVERSITY OF NORWAY • UIT •

https://en.uit.no/enhet/ifi

Thank You

Arctic LLM Workshop 2023
Dept. of Computer Science

bioai

www.bioailab.org