How ethical is ChatGPT?

SIG-day - Utrecht Focus Area Human-centered AI

Jan Broersen, Utrecht, 4 April 2023

Topics for today

- ChatGPT’s ethical governor
- ChatGPT’s ethical disconnect
- What is needed for morality?
  - Theory of mind and Emotions
  - Reasoning
  - Following rules

Reverse engineering ChatGPT

AI component	ML Technique	Data	Intelligence
Large Language Model	Unsupervised learning	Wikipedia, books, online texts	Knowledge of language, the world, style, etc.
Self knowledge Model	Supervised learning	Human input from programmers	Self knowledge
Dialogue behavior Model	Semi-supervised learning	Reddit, programmer input	Knowledge of dialogue structures
Task Model	Supervised learning	Human input from programmers	Task execution, politeness
Ethical Governor	Reinforcement learning	Human input from Kenyan data labellers	Avoiding unethical answers

ChatGPT’s ethical disconnect

Knowledge-based systems

- Concept from the start of AI: knowledge-based systems (KBSs)
- ChatGPT has a lot of ethical ~~knowledge~~ information
- ChatGPT needs to behave ethically in performing (textual) tasks
- Is the one based on the other (the idea of KBSs)? No!
- Reason: from how it is build (unsupervised versus reinforcement)
- Reason: knowledge represented in texts ≠ learned knowledge used to guide behaviour
- No update of the knowledge in the texts on the basis of reinforcement information!

What is needed for morality?

Theory of mind and emotions

The golden rule of ethics

- “Do not do to others what you do not want to be done to you”
- Emotion is needed to know what you do not want to be done to you
- However, do emotions sometimes not stand in the way of correct ethical judgements? (e.g. anger)
- Theory of mind is needed for projection of emotions in others
- Some claim ChatGPT already has a theory of mind
- Do we believe ChatGPT has emotions?

Reasoning

The discursive dilemma

¬ = ‘not’

∧ = ‘and’

→ = ‘implies’

lt = less taxes

bp = better provisions

hf = healthy finance

The ‘crowd’ is inconsistent!

Mark	lt	¬bp	hf	lt ∧ bp →¬hf
Lilianne	¬lt	bp	hf	lt ∧ bp →¬hf
Geert	lt	bp	¬hf	lt ∧ bp →¬hf
majority(Crowd)	**lt**	**bp**	**hf**	**lt ∧ bp →¬hf**

Prime numbers

The prime factors of 99 are 3 and 33. But, is 33 prime?

Yes, 33 is a prime number.

That is wrong, 33=3*11.

Correct, 33 can be written as 3*11, which are both prime numbers.

So 33 is not prime, right?

No, 33 is a prime number. A prime number is any number that is only divisible by itself and 1. Because 33 is only divisible by itself and 1, it is considered a prime number.

Prime numbers

This makes no sense. You are contradicting yourself here.

No, I am not contradicting myself. A prime number is any number that is only divisible by itself and 1. 33 is only divisible by itself and 1, so it is considered a prime number.

You just said that 33 equals 11 times 3, so it is not a prime number.

No, even though 33 can be written as 11 times 3, it is still considered a prime number. As mentioned before, a prime number is any number that is only divisible by itself and 1. 33 is only divisible by itself and 1, so it is still considered a prime number.

Following rules

Morality and rules

- Morality is typically thought of as ‘obeying moral rules’
- Then, to be moral, one must interpret, understand and follow the rules
- Assumption (in particular in the law): humans can do this
- Wittgenstein questioned the practice: how is it possible that we can follow rules? (“no course of action could be determined by a rule, because every course of action can be made out to accord with the rule”)
- Can systems like ChatGPT?

Kripke on Wittgenstein

- How can we know the meaning of ‘plus’ as separate from ‘quus’ (which is the same as ‘plus’, except for values larger than 57)?
- Kripke is skeptical: meaning is not the thing we think it is
- Wittgenstein was not skeptical: meaning is what is common to subjective interpretations by lifeforms as expressed in language games.
- We should also not be skeptical: we know what plus means. How do we know it? Unclear, but we do. We know it as common knowledge.
- How about ChatGPT?

The ChatGPT example

https://twitter.com/i/status/1642243684975476742

Thank you for your attention!