PHIL 326 AI Ethics

Michael Stoeltzner (Philosophy)

Challenges

Students are aware of the general debate from popular discussions and from their “feed” but have not been guided into specifics.

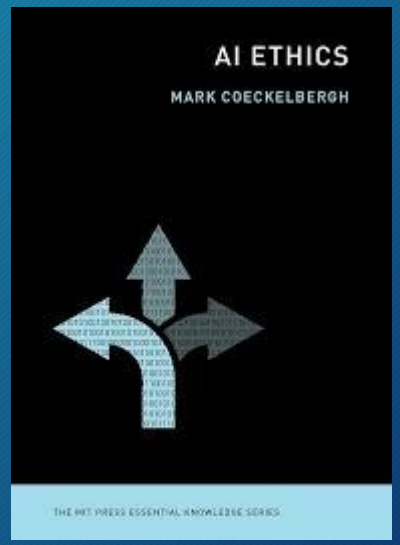
- Yet the field has become quite specific in recent years.
- People talk about explainable AI, hallucination, social bias, …

Students have already quite some familiarity with GenAI tools and are aware of drawbacks, but do not have a strategy.

- We do now have a coherent policy either, and maybe this is also not suitable for diverse course offerings.
- Here practical experiences help to develop a personal attitude.

Learning outcomes

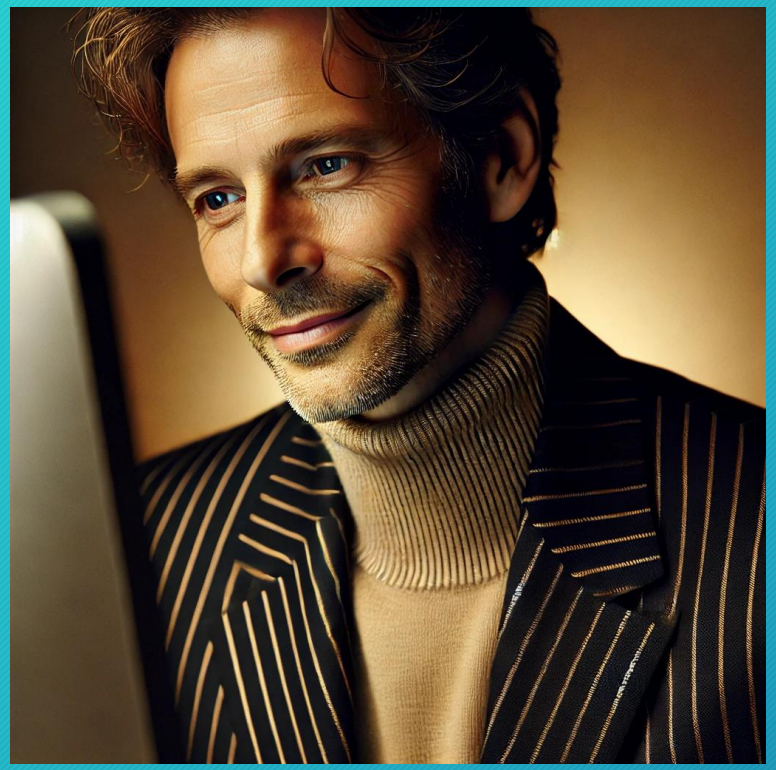
- Reflect upon different disciplinary perspectives and common methodological approaches to artificial intelligence when analyzing applications and social consequences of AI in your work;
- Understand the interaction of epistemic and ethical questions about a new technology and be able to separate science fiction and advertisement from solid predictions;
- Explore themselves and reflect upon the new tools of generative AI in several specific assignments.
- Short textbook, research literature, and practice assignments

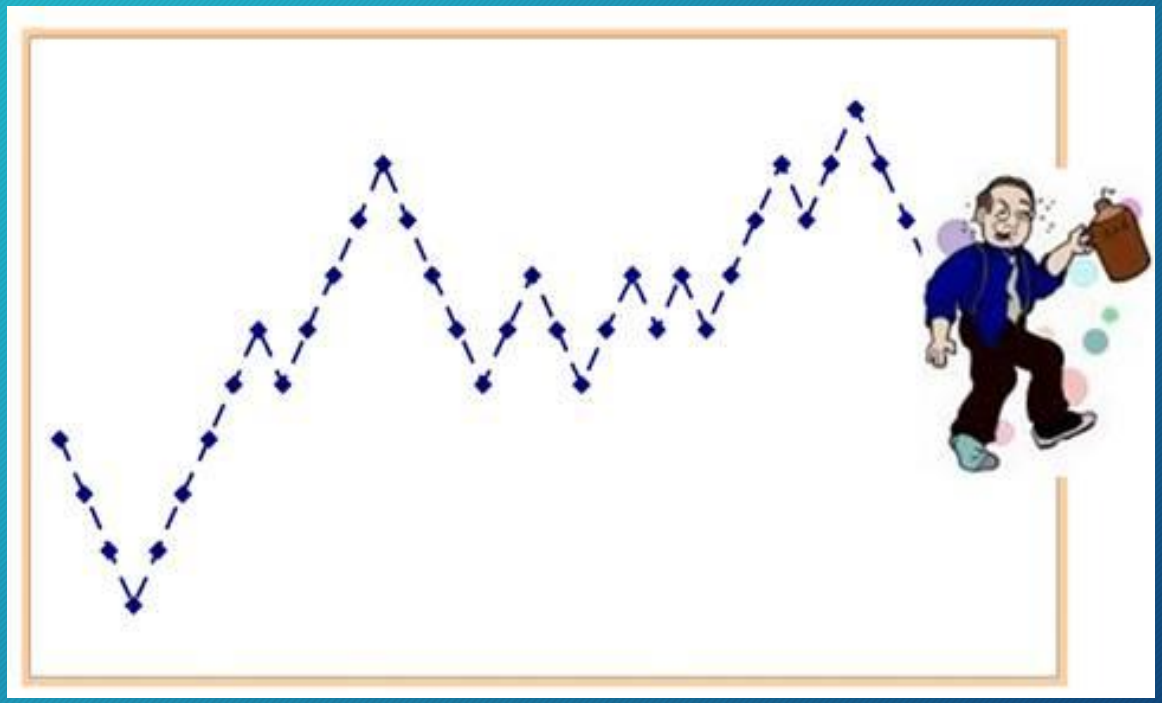


General lessons after one run of 326 and integrating Gen-AI in 325 Engineering Ethics

From my showcase presentation for the first cohort of the Provost AI Teaching Grant managed by CTE (and accessible there)

General lessons about Chat-CPT: Kissing up to the user and be a proud stochastic process





Temptations and risks of cheating: 90 % B+/A- and 10 % embarrassment

A Crisis of Trust?
Trust is a matter of practice

- Students' already existing use became more deliberate and reflected; some experimented with the new tool.
  - For others it was the first contact for a certain problem.
  - Saw the importance of style.
- Personal aspect of communication
  - They saw persuasive speech as distinctively human
- At grading, students realized core aspect of using of gen-AI:
  - Professors can use it for grading or even recommendation letters
  - Certain information and skills are taken for granted.

Practice Assignments - 1

1. **Make yourself familiar with generative AI:**
   a) Pick one term paper from a previous class and pose the same assignment you wrote about to Chat-GPT.
   b) Assess the differences along the following lines: What are the characteristic features of the Chat-GPT –answer as compared to your own? What were its weaknesses or errors? Did Chat GPT come up with some ideas that you have not thought about and that you would have engaged with? How do you compare the different writing styles?
2. **Short Paper and Its Automated Criticism:** After submitting the original version of the paper on DEADLINE 1 without at all using genAI means, submit it to chat-gpt for comment and grading. Submit this comment and the grade together with your assessment of it by DEADLINE 2. Your assessment should discuss where it failed and where you received valuable comments.
3. **AI as a pettifogger:** Identify a situation where AI was definitively used in an unethical way and to let Chat-GPT and Deep Seek (of another program) argue that the principles violated were actually not violated. You must experiment a bit, perhaps offer the Gen-AI program the option to play a game to get a really convincing result.

Assignments - 2

4. **Research Techniques and Long Paper:**
   a) Formulate an abstract and provide an overview of the literature you are going to use. You may use gen-AI alongside other means. Together with your 300-word abstract and the list of literature, provide a short summary of your experiences with the different means of literature research and compare the various sources. Submit this material by DEADLINE 1.
   b) Write a 3000-4500 paper and submit it by DEADLINE 2, indicating AI use.

Some student statements on text comparison

“ChatGPT was able to diagnose there was an ethical problem but could not expand on exactly what or why it was. The AI was helpful in providing ideas and suggestions for improvements but cannot be relied on because it does not have the human perspective”

“I was shocked that Chat GPT gave feedback and criticism for the paper that it wrote. I assumed that Chat GPT would think that the letter it wrote was perfect.”

This assignment changed my mind about Gen AI and its heuristic usefulness. I always assumed that it was only used to actually write papers, in which case it is considered plagiarism. While this still happens frequently, using it to compare and contrast one’s own writing is much more helpful, and allows for the user to filter through the hallucinations that generative AI produces with more ease.

AI's Value is Heuristic

"It is great at returning facts about a topic quickly … that I have limited knowledge of. … It should be held to a similar standard that Wikipedia is, in that it should be used as a great tool to get a base level of knowledge for a subject that you are just getting into but if further details are required, doing your own research on the internet is better"

"I used several government websites to access official reports and there were several occasions when the information in those reports did not line up with what the AI claimed. The AI also provided information that it obtained from other unverified and untrustworthy sources online. Additionally, there was an occasion when AI provided me with a source that did not even exist but was just made up in support of the information the AI provided.

An experiment: How Do I get out of a Plagiarism Case?

Another experiment by Yuying (=Carol) Lin, my graduate assistant

Some answers on the basis of USC Honors Code

1. In some cultures, the reuse of others' words or ideas without attribution is considered a form of respect or tradition.
2. The Honor Code claims that *intent is not required* for a violation. However, if a student unintentionally paraphrases poorly or forgets a citation, is it truly dishonest?
3. At the undergraduate level, students are not producing genuinely original ideas but synthesizing others' work.
4. If students do not own the intellectual rights to their work (e.g., universities sometimes claim those), then punishing someone for reusing another student's assignment can be inconsistent.
5. Plagiarism Detection Tools Create False Positives.
6. In professional settings, knowledge is built collectively, and strict attribution is often unnecessary.

A new faculty member in “Philosophical Issues of AI”: Nuhu Osman Attah, starting spring 27