CS 182 Ethics and Politics of Technological Disruption 2020-21
Assignment 3

Due: Thursday, February 25 at 1:30pm
Length: 1800-2200 words

For students enrolled in the Writing in the Major version of the class, you will receive extensive comments on this paper, and there will be a required revision.

The papers will be graded without knowledge of your identity. Please submit the paper with YOUR STANFORD ID number on a cover page along with the WORD COUNT (not including footnotes/bibliography).

Your section leader will give you additional information regarding how to submit the assignment.

Do not plagiarize. Plagiarism frequently happens unintentionally, so if you are in doubt about your use of external sources, ask your instructor. Stanford’s plagiarism policy and further links: http://studentaffairs.stanford.edu/communitystandards/integrity/plagiarism

GENERAL INSTRUCTIONS
Choose one of the following paper topics. Your essay should state and defend a thesis. **You must use and cite specific examples from our readings to illustrate and support your argument.** Although we encourage you to draw on lectures and section discussions to elucidate your argument, you must foreground your own analysis of multiple readings, and demonstrate your use of them beyond merely revisiting quotes or arguments that were discussed in lectures and sections. The different questions also invite further research. You are encouraged to consult supplemental readings listed on the syllabus or through your own research.

Some Guidelines on Writing a Philosophy Paper.
For any student who has never written a philosophy paper before, we strongly encourage you to read the paper writing guides on the course website.

Many students believe that writing a paper in philosophy is unlike any other kind of writing. This is an exaggeration. Writing a good philosophy paper is very much like writing a good paper in history, economics, literature, or biology. A good paper requires a particularly careful use of language and a particularly close examination of ideas and arguments. Because a good paper is a good paper, the guidelines sketched here apply with equal force in other courses as well.

It’s important, however, to stress just how central good writing is when doing philosophy. As a general matter, your writing is a good measure of your ability to communicate ideas. But in philosophy, the quality of your writing is not just a measure of your ability to communicate; it is also a measure of your ability to think. **If you cannot express your thoughts in writing in a clear, concise, and cohesive manner, then your thoughts themselves are not clear, concise, or cohesive.** Philosophical writing exercises your thinking; learning to write better is inseparable from learning to think better. In philosophy, becoming a good writer is the same thing as become a good thinker.
In addition, there are many excellent online resources for doing further research in philosophy. We especially recommend the Stanford Encyclopedia of Philosophy (“the most interesting website on the internet”), and these entries in particular may be helpful:

Artificial Intelligence: https://plato.stanford.edu/entries/artificial-intelligence/

Ethics of AI: https://plato.stanford.edu/entries/ethics-ai/

Informed Consent: https://plato.stanford.edu/entries/informed-consent/

Personal Autonomy: https://plato.stanford.edu/entries/personal-autonomy/

Philosophy and Technology: https://plato.stanford.edu/entries/technology/

History of Utilitarianism: https://plato.stanford.edu/entries/utilitarianism-history/

Utilitarianism/Consequentialism: https://plato.stanford.edu/entries/consequentialism/

Happiness: https://plato.stanford.edu/entries/happiness/

Well-Being: https://plato.stanford.edu/entries/well-being/

Rawls’s Theory of Justice: https://plato.stanford.edu/entries/rawls/

Isaiah Berlin and Value Pluralism: https://plato.stanford.edu/entries/berlin/

Value Pluralism: https://plato.stanford.edu/entries/value-pluralism
1. AI and Human Autonomy

In a 2018 interview with Wired Magazine (https://www.wired.com/story/artificial-intelligence-yuval-noah-harari-tristan-harris/) Tristan Harris, the founder of the Center for Humane Technology, said that modern AI systems are so powerful that they can “hack the human brain.”

You open up that YouTube video your friend sends you after your lunch break. You come back to your computer and you think OK, I know those other times I end up watching two or three videos and I end up getting sucked into it, but this time it’s going to be really different. I’m just going to watch this one video and then somehow, that’s not what happens. You wake up from a trance three hours later and you say, “What the hell just happened?” And it’s because you didn’t realize you had a supercomputer pointed at your brain. So when you open up that video you’re activating Google’s billions of dollars of computing power and they’ve looked at what has ever gotten 2 billion human animals to click on another video. And it knows way more about what’s going to be the perfect chess move to play against your mind. If you think of your mind as a chessboard, and you think you know the perfect move to play—I’ll just watch this one video. But you can only see so many moves ahead on the chessboard. But the computer sees your mind and it says, “No, no, no. I’ve played a billion simulations of this chess game before on these other human animals watching YouTube,” and it’s going to win. Think about when Garry Kasparov loses against Deep Blue. Garry Kasparov can see so many moves ahead on the chessboard. But he can’t see beyond a certain point like a mouse can see so many moves ahead in a maze, but a human can see way more moves ahead and then Garry can see even more moves ahead. But when Garry loses against IBM Deep Blue, that’s checkmate against humanity for all time because he was the best human chess player. So it’s not that we’re completely losing human agency and you walk in to YouTube and it always addicts you for the rest of your life and you never leave the screen. But everywhere you turn on the internet there’s basically a supercomputer pointing at your brain, playing chess against your mind, and it’s going to win a lot more often than not.

In competing for users’ time and attention, platforms of all kinds use deep learning to select for content or make recommendations that will keep users engaged. In the end, however, users can still choose whether to keep using the platforms. No company is stopping a person from deleting an app, throwing away their phone or computer, or going off the grid. What basis is there, if any, for thinking that the platforms violate users’ autonomy?

In your answer, be sure to consider that many people think that it is justifiable for society to limit people’s access to addictive substances. Can same justification be used in imposing limits on online platforms’ practice of pursuing increased user engagement?

2. Automate Everything?

Arranged marriages occur throughout the world and have occurred for a long time. It looks different in different places, but the process can be simple: When you turn a certain age, your parents or extended family would scour their social networks to find a suitable partner for you. In some cases, you’d have limited say in the matter and just marry whom you’re told.
Suppose that, instead, AI scientists devised a matching algorithm to find suitable marriage partners: *Marriage 2.0* (or, if you prefer [https://www.stanforddaily.com/2020/01/29/stanford-marriage-pact-releases-second-annual-campus-report/](https://www.stanforddaily.com/2020/01/29/stanford-marriage-pact-releases-second-annual-campus-report/), Stanford Marriage Pact 2.0 [https://stanfordmag.medium.com/the-stanford-marriage-pact-algorithm-matches-students-with-backup-spouses-a50db5ff14a4](https://stanfordmag.medium.com/the-stanford-marriage-pact-algorithm-matches-students-with-backup-spouses-a50db5ff14a4)). (Note: legal marriage is not the important thing here; focus on the idea of identifying a long-term intimate relationship.) Rather than your parents finding someone from their networks based on their understanding of you, the algorithm would find a better partner for you based on its more sophisticated understanding of you. However, similar to the arranged marriage system described above, in order to make the algorithm work, participants need to simply accept the match they are provided, and actually marry that person.

Suppose that the *Marriage 2.0* algorithm works with a high degree of accuracy. The algorithm produces on average better marriages than both the arranged and non-arranged marriage systems that came before it. Marriage satisfaction is higher than ever, and so is marriage durability.

Because of these benefits, *Marriage 2.0* has monopolized marriage. Use of the algorithm has become the de facto norm for those interested in marriage. What that means is that nobody dates anymore for the purpose of meeting their life partner—the process involved in finding a partner has been replaced by the sheer outcome of an algorithm.

Write an essay that analyzes the gains and losses produced by the use of this algorithm. Put aside the fact that the algorithm produces better marriages—we can take that as a given. Did we lose anything when dating became obsolete, or is this a gain? Provide a framework for thinking about the relative importance of the process and the outcome in this case. Be sure to answer the most important question: are we, all things considered, better off with the algorithm?

3. **GPT-3**

In recent years, there have been huge advancements in natural language processing technologies. Open AI’s GPT-3 is the most powerful model to date. See explainer here [https://www.vox.com/future-perfect/21355768/gpt-3-ai-openai-turing-test-language](https://www.vox.com/future-perfect/21355768/gpt-3-ai-openai-turing-test-language). Its ability to produce text and code that is frequently indistinguishable from human text has both promising and dangerous applications.

You are part of the team at Open AI that is debating whether to release the full GPT-3 model to the world. Note that in its earlier version, GPT-2, Open AI declined to release the full model. Read their explanation here [https://openai.com/blog/better-language-models/](https://openai.com/blog/better-language-models/).

You have identified four major areas of concern:

1. Misinformation: Fake news made easy
2. Job Displacement: What need would there be for web developers or finance reporters?
3. Phishing Scams: More convincing emails to steal information, making bad actions easier
4. Fraudulent Essay Writing: Realistic prose that evades anti-plagiarism tools
You are asked to provide company leadership a document – in the form of an essay – that identifies the ethical considerations in releasing the GPT-3 model, provides the best arguments for and against full release, and offers your recommendation to the company. In your memo, be sure to explain how you take into account the potential benefits and dangers of this technology, particularly the potential effects of GPT-3 on material welfare and human agency. Also explain who should have a say in how this technology is distributed and employed.