CSE199 Data, AI, and Society Week 2 Fall 2026
Individual Homework

AI Zoom: Boom or Doom?

No shortage of rhymes here—the title could be "AI Vroom: Bloom or Gloom?" or "AI Spume: Loom or Tomb?" This assignment calls for serious contemplation of challenging, yes horrid, scenarios. The goal of the **written homework** is to convey how they have moved from standard fare of fantasy sci-fi apocalypse to concrete particulars already impacting our world. The highlighted sources for this assignment break dangers into several areas and issues within each area, such that individual group members can choose and write on separate areas.

Two recent flumes of impact have been *AI-guided cyberattacks* represented by July's *HuggingFace* incident, and *AI's conquests in research-level mathematics and CS theory*. The first few lectures in the unit will discuss both. The latter may appear to matter less in the grand global picture, but (i) your instructor has ongoing direct experience with it, and (ii) mathematics has given rise to pervasive technology from atomic bombs to mass-market cryptography.

A key concept is **alignment**. This means the degree to which AI developments advance human welfare. The technical image is the angle between two vectors. The more the vectors point in the same direction, the more positive the cosine of the angle. But if they point away from each other, the cosine is negative and the effects are detrimental. This is a technical concept "under the hood" of training AI models themselves, but the idea matters at top level. We want AI outcomes to align well with—or at least not work against—as many "human value vectors" as possible. To see how complex this becomes, we need only think of mathematics: Advancing human knowledge by any means is innately a boon, but the "AI gold rush" has happened faster than the mathematical community has been able to adapt. A level-headed statement of health for our future with AI is, *Can we solve the alignment problem?*

Sources and Issue Categories

The **one required source** is an essay dated August 2026 by Dr. Xiaoyu He titled "Existential Risk From AI: An Exposition For Mathematicians." The essay is not just for mathematicians, but the author, in the Mathematics Department at Georgia Tech, is appealing first and foremost to his own professional community. The direct link is https://alkjash.github.io/ai-risk/, and a PDF mirror is in the course unit folder at https://cse.buffalo.edu/ regan/cse199/XiaoyuHeAIRiskAug2026.pdf. The essay's main virtue is its outline, which lists 20 statements to debate in five issue areas:

1. Immediate, Direct AI Takeover Scenarios.
2. Killer Tech Unlocked By AI (as if by magic).
3. AI Bootstraps Itself Into The Singularity.
4. Alignment Fails Intrinsically.
5. Alignment Fails Because of Human Failings.

One of the essay's major points is that multiple combinations of just one or a few of these can spell doom (the two lone math formulas given mainly serve this point). You, individually, need only be responsible for one of these five areas. Your group may collectively take multiple areas—see section "Continuation and Group Workflow" below.

For each area, you should take two or more statements that the essay puts within that area. Your task is to analyze the statements and find examples *in other sources* that pertain to them. A starter list of other sources is given below; you are welcome (indeed, encouraged) to seek and find others, but they must be cited.

Written Homework, due Week 2 Wed., 11:59pm

For the **individual written homework**, taking one area, it suffices to give **two examples**—apart from any mentioned in the essay itself—that illustrate statements in the area, and/or could bring them about. The examples can be drawn from other sources, including any in the list given below. The examples can be worrisome incidents (apart from the *Huggingface* attack itself) that have happened, or potential scenarios that have been seriously written about. For full credit, you should tie each example to one or more technical features of how Large Language Models work—as (will have been) covered in lectures—that contribute to bringing the danger about. A recommended format for a 1–2 page essay is to have one paragraph introducing the area, then two paragraphs giving and discussing your examples, and a conclusion giving your own estimation of the danger. (Did your stance change while doing this exercise?)

Your sources must be *published humanly written entities*, not AI chats. Note, however, the very bottom of Dr. Xu's essay: "Written with the help of Claude Fable 5 and OpenAI ChatGPT 5.6 Sol." Your written submission should include the sources for your examples as end citations.

Grading: Out of **3 pts.**, there is **1 pt.** for participation. Then **0.5** for choosing an issue area and describing it further, **0.5** for each of two examples for statements in that area, and **0.5** for the conclusion with your stance. Within each example, the final 0.2 is for giving one or more properties of LLMs—as covered in lectures—that in some way cause or enable the danger.

Workflow and Continuation Into Week 3

The Week 2 individual assignment does not presuppose any prior group interaction on its content. The one group interaction to agree on is to avoid all 3 or 4 members choosing the same one of the five issue areas. This is to ensure enough breadth for the continuation into the week-3 project, which is described in a separate document. It is fine for each member to choose a different area, and also fine for two members to choose the same area and even choose the same statements within the area. The week-3 presentation project will use **one** of each member's two examples, so even with such overlap, each member can have his/her own example to present. (Ah, having 3 choose the same area **and** choose the same two statements in that area also needs to be avoided, but 3 in one area, each on different statements, is AOK. See also "Freedom Note" below.)

Other than this one measure to avoid excessive overlap, the intent is that the written homework be done before any group consultation about the content of individual issues. The ideal is for each member to come individually versed in the chosen issues before group collaboration starts looking into week 3. This pacing and structure also aims to promote viewpoint diversity. We can't fully avert all group members coming to the same conclusions—which might be one member's voiced stance initially—but we can delay it if so. In sum, the week 2 homework really is intended as an individual assignment. See the separate Week 3 Project document for further group instructions.

Other Sources

This list may be expanded as the term progresses—indeed, the first several appeared after this term began. Not every source has examples of issues—some are just commentary. **Of course, you are welcome to seek and use other relevant published sources.** The sources listed here are a guide.

- Dario Amodei (CEO of Anthropic), "We Must Pace the Frontier." Post at darioamodei.com, 9/12/2026. Direct link: https://darioamodei.com/post/we-must-pace-the-frontier. (Some examples and concrete proposals.)

- Jakub Pachocki (Chief Scientist at OpenAI), "An Alien Mind." OpenAI website, 9/6/2026. Direct link: https://openai.com/index/an-alien-mind/. (Referred to as a go-slow warning, but IMHO less concrete than Amodei's essay.)

- Bill Gates, "The turbulent AI era is here. The choices we make now are critical." *Gates Notes* weblog, 8/26/2026. Direct link: https://www.gatesnotes.com/a-turbulent-ai-era-and-critical-choices-to-make. You are welcome, however, to read instead the "Cliff Notes" version at the ISOC Live Substack. There is also an interview of Gates about this by the MIT Technology Review.

  Gates highlights three dangers: the effect on jobs, empowerment of bad actors (a-la areas 2 and 5 by Dr. He), and the detriment of human relationships and capabilities.

My own earlier units in this course focused on the last—and you may add it as a permissible "area 6" to the five from the required essay.

- Tyler Cowen, "Should You Believe the Dire Warnings About 'AI Takeover'?" *The Free Press*, 8/30/2026. (Original link behind paywall, but mirrored here, and there is a related video. A commenter on X ascribed to Cowen the stance that AI safety creates as many opportunities for CS careers as AI takes away, but I do not find that specifically in this article. Disclaimer: Cowen is a personal friend from childhood.)

- James Dargan, "AI Safety Expert Roman Yampolskiy Believes AI Has a 99.9% Chance of Wiping Out Humanity," *AI Insider*, 7/6/2026. Free link. (Yampolskiy is often credited with coining the term "AI safety." He obtained his PhD at UB in 2008 under Professor Venu Govindaraju.)

- Andrew Critch and Jacob Tsimerman, "A Taxonomy of Omnicidal Futures Involving Artificial Intelligence," ArXiv https://arxiv.org/abs/2507.09369v1, 7/15/2025. (Much heavier going than Dr. He's essay. Tsimerman won one of this year's Fields Medals, then announced his departure from the University of Toronto's mathematics department to work on AI safety.)

- Liv Boeree,"The Dark Side of Competition in AI," TED talk posted 11/9/2023, direct link. Also see this recent interview. (I already showed parts of this video in my 2024 and 2025 units; it ties in to your "Prisoner's Dilemma" activity in week 1.)

- Richard Lipton and Kenneth Regan, "Human Extinction?" Weblog *Gödel's Lost Letter and P=NP*, 6/8/2023. https://rjlipton.com/2023/06/08/human-extinction/ (Well, I've been jolted off my own stance expressed there. You may ignore the technical second section.)

Freedom Note

This assignment description lays out structures and requirements for the sake of definiteness—and to make a rubric for grading that promotes even-handed treatment of students from various backgrounds and experiences. If, however, you happen to have some particular experience that is fully relevant, then you may deviate from strict adherence. For example, you may then focus more deeply on just one issue, with just one example made all the richer by your experience. The guidelines here are that (i) your writeup must make clear how you are adding at-least-equivalent value to what the above rubric prescribes, and (ii) you must integrate your experience with your group in a way that allows the others to "put their oars in," even as critics—or as potential later teammates. Even so on the latter, we do not want the entire group to be captured by the one issue—hence again the desire that at least one group member work in a different issue area.