Prof. McCormick
Department of Philosophy
MND 3020 CSUS
mccormick@csus.edu

Philosophy of AI Weekly Schedule.
Phil 192m, section 1, MW 12:00-1:15pm, DH 110

Readings and Lectures	Quizzes and Assignments	
Week of Aug. 25 Readings: Muller: Philosophy of AI, A Structured Overview Lee and Trott: LLMs Explained Optional background: Roser: A Brief History of AI Class Discussion Slides: AI First Week Intro, History, Terms, Philosophical Issues Assignment: Learn with an LLM Create Chat Account at ChatGPT Edu	**Dropping the Course:** *Students who wish to drop themselves through the second week should use the my.csus.edu portal and click on the OnBase Forms link.* Add/Drop/Withdrawal Guidelines **Success in College:** *How to do well in college courses.* *How to Succeed in a College Philosophy Course video* *How to correspond with your Professor* Learning from our Mistakes **Quiz:** Learn with an LLM in-class quiz on Wednesday.	
Week of Sept. 1 No class on Monday, Labor Day Class Discussion Slides:Intro, History, Terms, Philosophical Issues Schwitzgebel: Defining Artificial Intelligence Stufflebeam: Connectionism: An Introduction, part 1 Connectionism: An Introduction, part 2 Connectionism: An Introduction, part 3 McCormick Lecture: on Connectionism GOFAI vs. ANN	Wednesday: In-class reading quiz on Learn with an LLM from last week, and on Muller, Lee and Trott readings from last week.	

Week of Sept. 8 **Consciousness** Readings: Chalmers: Could a LLM be Conscious? Chalmers' slides: Are LLMs Conscious? Schwitzgebel: The Mimicry Argument Against Robot Consciousness Class Discussion Slides: AI Consciousness Optional background: Butlin, Long, Elmoznino, Bengio, Birch, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. Birhane, McGann, Large Language Models of what? Mistaking engineering achievement for human linguistic agency	In class reading quiz on Connectionism from last week and and Chalmers readings for this week	
Week of Sept. 15 **Consciousness** Readings: Chalmers: Could a LLM be Conscious? Chalmers' slides: Are LLMs Conscious? Schwitzgebel: The Mimicry Argument Against Robot Consciousness Class Discussion Slides: AI Consciousness Optional background: Butlin, Long, Elmoznino, Bengio, Birch, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. Birhane, McGann, Large Language Models of what? Mistaking engineering achievement for human linguistic agency	In class reading quiz on Schwitzgebel readings for this week.	

Week of Sept. 22 **Singularity** Singularity, Future of Humanity and AI Bostrom, Vulnerable World Take off scenarios, intelligence explosion, implications Class Discussion Slides:	Readings: Lectures:	
Week of Sept. 29 **Singularity** Singularity, Future of Humanity and AI Bostrom, Vulnerable World Take off scenarios, intelligence explosion, implications Class Discussion Slides: Eth, Davidson: Will AI R & D Automation Cause a Software Intelligence Explosion? Empirical evidence suggests that, if AI automates AI research, feedback loops could overcome diminishing returns, significantly accelerating AI progress Kokotajlo, Alexander, Larsen, Lifland, Dean. AI 2027 We predict that the impact of superhuman AI over the next decade will be enormous, exceeding that of the Industrial Revolution. We wrote a scenario that represents our best guess about what that might look like. It's informed by trend extrapolations, wargames, expert feedback, experience at OpenAI, and previous forecasting successes. Alexander, Kokotajlo (Patel). AI 2027: Month by Month Model of Intelligence Explosion podcast https://situational-awareness.ai/	Readings: Lectures:	

Bailey, Brynjolfsson, Korinek. Machines of mind: The case for an AI-powered productivity boom Large language models such as ChatGPT are emerging as powerful tools that not only make workers more productive but also increase the rate of innovation, laying the foundation for a significant acceleration in economic growth. As a general purpose technology, AI will impact a wide array of industries, prompting investments in new skills, transforming business processes, and altering the nature of work. However, official statistics will only partially capture the boost in productivity because the output of knowledge workers is difficult to measure. The rapid advances can have great benefits but may also lead to significant risks, so it is crucial to ensure that we steer progress in a direction that benefits all of society.		
Week of Oct. 6 **Allignment and Control Problem** Bostrom, recent piece? Gabriel, Iason. Artificial Intelligence, Values, and Allignment This paper looks at philosophical questions that arise in the context of AI alignment. It defends three propositions. First, normative and technical aspects of the AI alignment problem are interrelated, creating space for productive engagement between people working in both domains. Second, it is important to be clear about the goal of alignment. There are significant differences between AI that aligns with instructions, intentions, revealed preferences, ideal preferences, interests and values. A principle-based approach to AI alignment, which combines these elements in a systematic way, has considerable advantages in this context. Third, the central challenge for theorists is not to identify 'true' moral principles for AI; rather, it is to identify fair principles for alignment that receive reflective endorsement despite widespread variation in people's moral beliefs. The final part of the paper explores three ways in which fair principles for AI alignment could potentially be identified.	Readings: Lectures:	

Class Discussion Slides:		
Week of Oct. 13. *Fall Ethics Symposium* continued. Discussion Slides:	Readings: Lectures: Midterm Exam Questions:	
Week of Oct. 20 Sharing the world Bostrom and Shulman, Sharing the World with Digital Minds Bostrom and Shulman: Propositions Concerning Digital Minds and Society Background: AI Ethics at SEP *Midterm Review*		
Week of Oct. 27 **In class midterm.**	*Please adopt an every-other-seat pattern. No one will be admitted to the midterm after ten minutes into the class, and no one may leave the room until they finish. No phones, no hats, no earbuds or headphones, no smart watches. No makeup exams or reschedules; plan accordingly.* Readings: Lecture:	
Week of Nov. 3 **AI Rights** Schwitzgebel and Garza:A Defense of the Rights of Artificial Intelligences (pdf, September 15, 2015). Designing AI with Rights, Consciousness, Self-Respect, and Freedom Schwitzgebel, Eric. Against Designing "Safe" and "Aligned" AI Persons (Even if They're Happy) Slides:	Readings: Lecture:	
Week of Nov. 10		

AI Rights Schwitzgebel and Garza:A Defense of the Rights of Artificial Intelligences (pdf, September 15, 2015). Designing AI with Rights, Consciousness, Self-Respect, and Freedom Schwitzgebel, Eric. Against Designing "Safe" and "Aligned" AI Persons (Even if They're Happy) Slides:	Readings: Lecture:	
Week of Nov. 17 Killer robots Russell: Banning Lethal Autonomous Weapons: An Education Simpson and Muller: Just War and Robots' Killings Optional: Sparrow: Killer Robots Slides:	Readings Lecture:	
Week of Nov 24 Russell: Banning Lethal Autonomous Weapons: An Education Simpson and Muller: Just War and Robots' Killings Optional: Sparrow: Killer Robots Nyholm and Frank. It Loves Me, It Loves Me Not: Is It Morally Problematic to Design Sex Robots that appear to Love Their Owners? Slides:	Readings and lectures for next week: Lectures:	

Week of Dec. 1 Final Review: come prepared with your notes and questions. Slides: *Complete the online instructor review on Canvas by the end of the day Friday. If 80% or more of the class does it, everyone gets 2-5 additional point on the in-class points portion of the semester grade..*	Final exam review:	
Week of Dec. 8: *CSUS Finals week* Final Exam Review Sheet ***Final Exam in class,*** *Please adopt an every-other-seat pattern. No one will be admitted to the midterm after ten minutes into the class, and no one may leave the room until they finish. No phones, no hats, no earbuds or headphones, no smart watches. No makeup exams or reschedules; plan accordingly.* The Academic Honesty Policy, CSUS will be strictly enforced.		

History
Intro Overviews
    Connectionism
    LLMs, RL

Artificial Intelligence, Internet Encyclopedia of Philosophy

Ethics of AI IEP, Internet Encyclopedia of Philosophy

Ethics of AI SEP, Stanford Encyclopedia of Philosophy

Muller, Philosophy of AI: A Structured Overview

Muller, Ethics of AI and Robotics.

> Artificial intelligence (AI) and robotics are digital technologies that will be of major importance for the development of humanity in the near future. They have raised fundamental questions about what we should do with these systems, what the systems themselves should do, what risks they involve and how we can control these.
>
> After the Introduction to the field (1), the main themes of this article are: (2) Ethical issues that arise with AI systems as objects, i.e. tools made and used by humans; here, the main sections are privacy and manipulation, opacity and bias, human-robot interaction, employment, and the effects of autonomy. (3) AI systems as subjects, i.e. when ethics is for the AI systems themselves in machine ethics and artificial moral agency. (4) The problem of a possible future AI superintelligence leading to a 'Singularity'.
>
> For each section within these themes, we provide a general explanation of the ethical issues, we outline existing positions and arguments, then we analyse how this plays out with current technologies and finally what policy consequences may be drawn.

Lee, Trott. Large Language Models, explained with a minimum of jargon and math.

> We'll start by explaining word vectors, the surprising way language models represent and reason about language. Then we'll dive deep into the transformer, the basic building block for systems like ChatGPT. Finally, we'll explain how these models are trained and explore why good performance requires such phenomenally large quantities of data.

Brief History of AI

Stufflebeam: Connectionism: An Introduction, part 1

Connectionism: An Introduction, part 2

Connectionism: An Introduction, part 3
McCormick Lecture: on Connectionism

Topics
    Consciousness, Thinking
        Butlin

Butlin, Long, Elmoznino, Bengio, Birch, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.

> Whether current or near-term AI systems could be conscious is a topic of scientific interest and increasing public concern. This report argues for, and exemplifies, a rigorous and empirically grounded approach to AI consciousness: assessing existing AI systems in detail, in light of our best-supported neuroscientific theories of consciousness. We survey several prominent scientific theories of consciousness, including recurrent processing theory, global workspace theory, higher-order theories, predictive processing, and attention schema theory. From these theories we derive "indicator properties" of consciousness, elucidated in computational terms that allow us to assess AI systems for these properties. We use these indicator properties to assess several recent AI systems, and we discuss how future systems might implement them. Our analysis suggests that no current AI systems are conscious, but also shows that there are no obvious barriers to building conscious AI systems.

Chalmers

Objections: Birhane, Schwitzgebel

Chalmers, David. Could a Large Language Model be Conscious?

> My conclusion is that within the next decade, even if we don't have human-level artificial general intelligence, we may well have systems that are serious candidates for consciousness. There are many challenges on the path to consciousness in machine learning systems, but meeting those challenges yields a possible research program toward conscious AI.

Slides: Could a Large Language Model be Conscious?
Lecture: Are LLMs Sentient lecture at NYU

Turing Test

Singularity, Future of Humanity and AI

Bostrom, Vulnerable World

Allignment Problem

Bostrom: Superintelligent Will

Shulman, Bostrom: Sharing, and Propositions

Moral Status of AI

Morality of Building AI