STS 351 Minds and Machines

Instructor: Dr. Daniel Estrada
E-mail: estrada@njit.edu
Zoom Meeting Link

Office: Cullimore 419
Office Hours: T 1-2pm and by appt.
Discord: discord.gg/NxFvdH7

Class Meeting: F 8:30-11:30 KUPF 107

Course Description: This course introduces a variety of key concepts, themes, and historical debates in the philosophy of minds and machines. The course is particularly concerned with the proposal that the mind is a computing machine, and the sociotechnical and ethical consequences of such views. In Unit 1 we will cover the history of computing machinery and intelligence, including classical logic and computation, and current machine learning techniques. In Unit 2 we will look at specific issues in the philosophy of AI, including perception, autonomy, and understanding.

Prerequisites: STS 201 and STS 210, each with a grade of C or better.

Course objectives:
- Develop an understanding of key issues in historical and contemporary debates in the philosophy of AI.
- Understand various arguments for and against a computational theory of mind, and the historical, scientific, and technological context in which they arise.
- Engage in peer-led debate and discussion on critical issues like agency, consciousness, and machine learning.
- Conduct scholarly research through short peer reviewed essays
- Prepare group presentations and conversations exploring the central themes of the course.

Lesson Plan

Unit 1: Thinking machines
       Lesson 1: History of AI and computation
       Lesson 2: Logic and Computation
       Lesson 3: Turing Machines & GOFAI
       Lesson 4: Neural networks and machine learning
       Lesson 5: Turing and Lovelace
Unit 2: Mechanisms of mind
       Lesson 6: Mechanism and organization
       Lesson 7: Life and mind
       Lesson 8: Sentience and consciousness
       Lesson 9: Understanding
Unit 3: AI Ethics

Lesson 10: Algorithmic bias
Lesson 11: Autonomous weapons and vehicles
Lesson 12: Creativity and copyright
Lesson 13: Robot rights
Lesson 14: Ethical Treatment of Robots
Lesson 15: Course wrap up

Assignments and expectations

- **Attendance:** Regular classroom attendance is **required**. (19%)
- **Presentation**: One 10-15 min presentation with slides on readings given in class Schedule your presentation on Canvas. (10%)
- **Reading notes**: 300+ words of "reading notes" due for each lesson reflecting on reading assignments. Must be submitted before class for full credit. (28%)
- **Podcasts:** Students are required to participate in a 1 hour recorded group conversation at the end of Unit 1, 2, and 3. (22.5%)
- **Encyclopedia Entry:** Students are asked to complete one 3-5 page discussing one important concept in the philosophy of mind and AI. This assignment will require some scholarly research and analysis. (5%)
- **Final Paper:** Students are required to complete one 7-10 page essay on a topic of their choosing. Preparation will involve a drafting round and a meeting with the instructor. (15%)

Grade break down

- Attendance and Participation = 13 x 15pts = 195 (19.5%)
- Presentation = 100 (10%)
- Notes 14 x 20pts = 280 (28%)
- Podcasts 3 x 75 pts = 225 (22.5%)
- Encyclopedia Entry = 50 pts (5%)
- Final paper = 150 pts (15%)
    - Final paper = 75
    - Final paper proposal = 5
    - Final paper draft = 20
    - Final paper meeting = 50

Total = 1000 pts

Grade Scale
Final grades are calculated on the following scale:

A: 900+
B+: 850+

B: 800+
C+: 750+
C: 650+
D: 500+
F: < 500

There is a 5 point tolerance for bumping a grade to the next letter when calculating final grades.

Assignment Schedule

Reading notes are due before class starts according to the schedule below. All assignments must be submitted to the appropriate place on Canvas. See the syllabus for details on assignments and class expectations.

Unit 1: Thinking machines
Lesson 1: History of AI and computation
F 1/24 **Introductions**
   Presentation Schedule Due
   L1 Notes

Lesson 2: Logic and Computation
F 1/31 **L2 Notes**

Lesson 3: Turing Machines & GOFAI
F 2/7 **L3 Notes**

Lesson 4: Neural networks and machine learning
F 2/14 **L4 Notes**

Lesson 5: Turing and Lovelace
F 2/21 **L5 Notes**
   U1 Plagiarism Check

Unit 2: Mechanisms of mind
 Lesson 6: Mechanism and organization
F 2/28 **L6 Notes**
F 2/28 **Unit 1 Podcast due**
   Encyclopedia Entry Proposals due

Lesson 7: Life and mind
F 3/7 **L7 Notes**

Lesson 8: Sentience and consciousness
F 3/14 **L8 Notes**
      Encyclopedia Entry due

SPRING BREAK

Lesson 9: Understanding
F 3/28 **L9 Notes**
      U2 Plagiarism check

Unit 3: AI Ethics
Lesson 10: Algorithmic bias
F 4/4 **L10 Notes**
      U2 Podcast Due

Lesson 11: Autonomous weapons and vehicles
M 4/7 LAST DAY TO WITHDRAW
 F 4/11 **L11 Notes**

Lesson 12: Creativity and copyright
F 4/18 **L12 Notes**
      Final Paper proposal

Lesson 13: Robot rights
F 4/25 **L13 Notes**
      Final paper meeting

Lesson 14: Ethical treatment of robots
F 5/2 **L14 Notes**
      Final paper meeting

Lesson 15: Course wrap up
W 5/7 **Final Paper peer review in class**
F 5/9 **U3 Podcast due**
      U3 Plagiarism Check

W 5/14 **Final Paper Due**

Su 5/18 **Grades Due**

Assignment Details

Attendance: Regular class attendance is required, and earns up to 100 points of credit for the semester. Students can miss up to three class sessions before an impact on attendance grade. There are 28 total days of class, so 25 attendance days earn full credit. On time attendance counts for one day. Attendance is considered late if registered more than 5 minutes after class begins and earns 80% credit. Attendance is taken on the class Discord server. Please do not register attendance on Discord until you are actually in your seat in class. Students registering attendance without being in class will lose all attendance credit for the semester. Note that attendance for some classes also earns participation credit! See more info below.

Participation: Participation credit is earned by participating in some classroom activities. This includes the Introduction thread at the start of the semester, see instructions on Canvas. Classroom attendance and participation in these activities will earn participation credit. If you cannot attend class on these days, let me know and I will offer an alternative assignment to make up this participation credit.

Presentations: Students must prepare a 10-15 minute presentation on one of the readings in class. Slides are encouraged but not required. The presentation should offer a close reading of the text, summarizing and explaining (in the student's own words) the main conclusions, concepts, and perspectives discussed in the article. Students are encouraged to develop a critical reading of the text by (for instance) drawing out associations with other class readings and discussions, contributing independent research on the same topic, or offering their own critical analysis and insights on the paper and topic. However, the primary focus of the presentation should be on elucidating the readings for class discussion. Students can work individually or in groups. Presentations must engage the primary readings, but they can also engage with supplemental readings and independent research. Basically, your presentation is your primary opportunity to lead the classroom discussion through the reading list. Take advantage of it!

Reading Notes: Students are expected to complete 300+ words of reading notes each week, posted on Canvas in the appropriate discussion thread. Reading notes document student engagement with the weekly readings. Notes can engage either required or supplemental readings. Notes don't need to be structured as a formal essay. Scattered thoughts and reactions, bullet points, sketches of ideas, etc are fine. However, notes should be primarily in your own words. Notes should not consist mostly of quotes or paraphrasing from the source material. You can include quotes you find important or interesting in your notes, but you should also explicitly explain and react to the quote in your own words. The quote itself doesn't contribute to your notes word count. Notes will be scrutinized for plagiarism, so please be careful to write your notes in your own words! Notes submitted before class Wednesday earn up to 20/20 points. Notes submitted by Friday earn up to 15/20 points. Notes submitted after one week late earn a maximum of 10/20 points.

Papers are scholarly essays substantively engaging with the debates and ideas found in the readings and lectures for each unit of the course. Papers should demonstrate a clear understanding of the issues presented in the readings and careful critical analysis of the scholarly texts. Papers can be argumentative and defend a particular position or controversial thesis in the debate. Papers can also be clarificatory, seeking to elucidate some complex issue or concept through additional research and reflection. Papers will be developed over several activities at the end of the semester, during which proposals and drafts will be made available for peer review and feedback. Detailed instructions and schedules are on Canvas.

Honors Requirements: Students registered for the honors section of the course are required to complete all regular assignments, with the following additions:
- Regular attendance is required for honors students
- 16 Reading notes required (up from 12)
- Honors students must prepare 2 presentations (up from 1).
- Honors students must schedule a meeting with the instructor to discuss the U1 paper
- Honors students must complete one of the following:
    - Either write a short (3-5 page) U2 paper (Can be an Encyclopedia write up)
    - Or write 2 additional reviews to U1 papers
    - Or U3 paper must be 10-15 pages (up from 5-7)

Accessibility policy: I want all students to succeed in this class, and I will gladly accommodate the special circumstances and needs of all students to make sure that happens. I understand that life doesn't happen on the semester schedule, and that school work can't always be a top priority. In pandemic conditions we all need to be more flexible with scheduling and difficult work conditions; I understand how medical issues or disability can complicate these challenges. If there is any issue impacting your performance in class, please come talk to me in office hours or send me a message by email or on Canvas! Even if you're behind on assignments, drop me a message letting me know what's up, I'm sure we can figure something out =)

Late policy: Assignments earn a small late penalty for material submitted after the assignment due dates posted on Canvas. I'll allow a short (~30 min) grace period for assignments due at midnight; assignments received at 12:01am will not be marked as late, but assignments received at 2am will. Late assignments are accepted until the end of the Unit.

Excused Absences: If you have a legitimate excuse that you know about in advance (an academic conference, athletic event, National Guard duty, expected delivery date, etc.), please make arrangements with me in advance. Extensions for anticipated issues must be arranged at least 48 hours before a deadline to avoid a late penalty. Unexpected emergencies (medical emergencies, deaths in the family, etc.) should be brought to the attention of the Dean of Students with the Student Concern Reporting Form. The Dean's office is equipped to verify your situation confidentially and provide the administrative support you need. The Dean's office can also coordinate with all your instructors for any issues that arise. After an emergency and when you are able to return to school work, let me know what's up (a short note will do). I'll

recommend you contact the Dean with the form linked above if you haven't already, and we can discuss a plan for completing your missing work, and go from there.

Plagiarism Policy

[Plagiarism Slides](#)

Plagiarism means using work that you did not produce, but presenting it as if it is your own work in assignments. If you did not write the words yourself, you must clearly distinguish that work from your own with quotes and citations. When I am scanning for plagiarism I am looking for long blocks of text that are clearly taken from other sources (possibly with minor modifications) without proper attribution and without distinguishing it from the students own work. Passing off the work of another for credit is plagiarism, and it will not be tolerated in an ethics course.

Copying and pasting from the web is a form of plagiarism*. **Changing a few words in an extensively quoted passage is a form of plagiarism. Using AI text generators like chatGPT is a form of plagiarism. Failing to provide adequate citations is a form of plagiarism. Copying from your own work (including work from previous semesters) without acknowledgement counts as plagiarism.** In general, you should never copy large blocks of text from any other source and present it in your own essay as if it were your own words. That includes copying text from online text generators or language translators. Check* [*this link*](#) *for a detailed explanation of legitimate paraphrase and illegitimate plagiarism. Any work you use should be given adequate citation so your readers can find and review your sources. Just as in mathematics, you need to show your work! If you use any source in your research, (including dictionaries, Wikipedia and other encyclopedias, and translation tools) even if you don't quote it directly, provide a citation.*

To avoid plagiarism, you must clearly distinguish your work from the work of others. Any work taken from others must be identified with "quotation marks" and explicit citation. Changing a few words in a quote does not make it your work. *If you use online text generators (like chatGPT, Grammarly, or other text sources), you must explicitly identify that text as not being your own work. You must also cite the explicit generator used, including the version and dates it was used. If you use AI text generators at all, you must also supply the full prompt history generating that text as an appendix to your assignment. If you wrote the essay in another language and then used a translator, you should provide the original text in the original language with your submission. If you read a script in any presentation, you must include the text of that script to the plagiarism detection software on Canvas. If you translate your essay from another language, you must include the original untranslated text for comparison. Failure to do so will not earn credit.*

Suspected cases of plagiarism will be given zero credit for the assignment with a warning about the plagiarism policy. Students found plagiarizing will also forfeit all extra credit opportunities for the semester. *Repeated or extreme instances of plagiarism will be reported directly to the Dean of Students as a violation of the [Student Code of Academic Integrity](#)* **Note:** *the research project is a honeypot for cheaters, and typically results in multiple instances of plagiarism in each section. I won't hesitate to fail students who cheat in my ethics course. Consider this your first warning.*

I have substantially reorganized my class around group discussions and presentations to discourage the use of AI text generators. None of the writing assignments in class are "busy work". They all ask you to demonstrate direct engagement with the readings and with the ideas and perspectives of your fellow students. Please take this opportunity to engage your peers in discussions on ethics seriously!

See these [Plagiarism Slides](#) with detailed information on the NJIT and course policies on plagiarism, including examples of legitimate and illegitimate paraphrase, to help you understand the plagiarism policy.

NJIT Plagiarism Policy

"Academic Integrity is the cornerstone of higher education and is central to the ideals of this course and the university. Cheating is strictly prohibited and devalues the degree that you are working on. As a member of the NJIT community, it is your responsibility to protect your educational investment by knowing and following the academic code of integrity policy that is found at:

[*http://www5.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf*](http://www5.njit.edu/policies/sites/policies/files/academic-integrity-code.pdf)

Please note that it is my professional obligation and responsibility to report any academic misconduct to the Dean of Students Office. ***Any student found in violation of the code by cheating, plagiarizing or using any online software inappropriately will result in disciplinary action. This may include a failing grade of F, and/or suspension or dismissal from the university.*** *If you have any questions about the code of Academic Integrity, please contact the Dean of Students Office at dos@njit.edu"*

Readings

Check Canvas for an updated reading list

Unit 1: Thinking Machines
Lesson 1: History of Automation

Estrada (2023) History of AI audio lecture and slides

- Mullaney et al (2021) Your Computer on Fire (Intro, Ch 1, 2, 6, 7, 8, 9)
- Benjamin (2019) Race after technology Intro, Ch 2, Ch 3
- BobbyBroccoli (2022, YouTube) The image you can't submit to journals anymore

Lesson 2: Logic and Computation

Estrada slides: Arguments, logic, and computation
Logic

- Schotch (2006) Introduction to logic and its philosophy
- Shapiro and Kissel (2022 SEP) Classical logic
- Parsons (2017, SEP) The traditional square of opposition
- Crash Course (2016, YouTube) Philosophical Reasoning

Computation

- Aaronson (2011) Why philosophers should care about computational complexity
- Chalmers (1994) On implementing a computation
- Piccinini (2021, SEP) Computation in physical systems
- Horst (1999) Symbols and computation
- Haugeland (1981) Semantic Engines: Introduction to Mind Design

Lesson 3: Turing Machines & GOFAI

Estrada slides: The day computers took over the world

Good Old Fashioned AI (GOFAI)

- Haugeland (1981) Semantic Engines: Introduction to Mind Design
  - Turing (1947) Lecture on the Automatic Computing Engine
  - Turing (1950) Computing Machinery and Intelligence
  - Saygin (2000) Turing's test, 50 years later
  - Searle (1980) Minds Brains and Programs

Computational Theory of Mind (CTM)

- van Rooij et al (2023) Reclaiming AI as a theoretical tool for cognitive science

- ○ Haugeland (1994) On the nature and plausibility of cognitivism
- ○ Piccinini (2010) Mind as neural software?
- ○ Chalmers (2012) Computational foundations of cognitive science
- ○ Gallistel (2017) The neurobiological bases for the computational theory of mind
- ○ Rescorla (2020, SEP) Computational theory of mind

Lesson 4: Neural networks and machine learning

Estrada slides: Machine learning and baby computers
Artificial neural networks

- ● Bender et al (2021) On the dangers of stochastic parrots
  - ○ Hinton (2007) The next generation of neural networks
  - ○ LeCun, Bengio, & Hinton (2015) Deep Learning
  - ○ Vaswani et al (2017) Attention is all you need
  - ○ Marcus (2018) Deep Learning: A critical appraisal
  - ○ Kriegeskorte & Golan (2019) Neural Network Models and Deep Learning

Explainers and demos

- ● Art of the Problem (2023): How neural networks learned to talk,
  - ○ How NNs learn
  - ○ How NNs learn concepts
- ● Tensorflow Playground (demo)
- ● Tensorflow Embedding Projector (demo)
- ● 3blue1brown: Neural Networks. video series (S3 E1-4)
- ● Computerphile: Neural Networks video series
  - ○ How AI image generators work
  - ○ Stable Diffusion in code
  - ○ How GPT3 works
  - ○ AI Language models and transformers

Lesson 5: Turing and Lovelace

- ● Turing (1950) Computing Machinery and Intelligence
  - ○ Saygin (2000) Turing's test, 50 years later
  - ○ Searle (1980) Minds Brains and Programs
- ● Natale and Hendrickson (2022) The Lovelace effect: Perceptions of creativity in machines
  - ○ Bringsjord (2000) Creativity, the Turing Test, and the (Better) Lovelace Test
  - ○ Abramson (2008) Turing's responses to two objections
  - ○ Whittaker (2023) Plantations, Computers, and Industrial Control

L6: Mechanisms, organization, and autonomy

Mechanical philosophy

- Craver (2015, SEP) Mechanisms in Science
- Mitchell (2009)Complexity, A guided tour
  - Keller (2008) Organisms, Machines, and Thunderstorms: A history of self-organization Part One, Part Two
- Levy and Bechtel (2013) Abstraction and the organization of mechanisms
  - Bechtel (2013) Thinking Dynamically about Biological Mechanisms
  - Bechtel, (2012) Why do Biologists use So Many Diagrams?

Agency and Autonomy

- Bechtel (2007) Biological Mechanisms: Organized to Maintain Autonomy
  - Bechtel and Abrahamsen, Mental mechanisms, autonomous systems, and moral agency
- Villalobos and Dewhurst (2018) Enactive autonomy in computational systems
- Estrada (2018) Conscious enactive computation
- Bandarian et al (2009): Defining Agency
- Moreno and Mossio (2015) Biological Autonomy
- Fischer et al., Four views on Free Will
- Frankfurt, Freedom of the Will and the Concept of a Person

Lesson 7: Life and mind

- Thompson (2007) Mind in life (precis)
- Mayr (2005) What makes biology unique?
- Godfrey-Smith (2016) Other Minds The Octopus, The Sea, and the Deep Origins of Consciousness
- Godfrey-Smith (2020) Metazoa Animal Life and the Birth of the Mind
- Godfrey-Smith (2023, Youtube) Whitehead lectures: Limits of sentience, Boundaries of consideration
- Smith (2015) New Theories on the Origin of Life
- Dennett (1996) Kinds of minds
- Brooks (1991) Intelligence without representation
- Cruse and Shilling (2015) From walking to consciousness
- Webb (2019) Insects as a cognitive edge case

Lesson 8: Sentience and Consciousness

- Nagel (1974) What is it like to be a bat?
- Schwitzgebel (2015) If materialism is true the united states is probably conscious
- Cruse and Shilling (2015) From walking to consciousness
- Graziano (2013) Consciousness and the social brain
- Godfrey-Smith (2016) Other Minds The Octopus, The Sea, and the Deep Origins of Consciousness
- Godfrey-Smith (2020) Metazoa Animal Life and the Birth of the Mind

- Godfrey-Smith (2023, Youtube) Whitehead lectures: Limits of sentience, Boundaries of consideration
- Estrada (2023) The AIdeal: sentience and ideology

Lesson 9: Understanding

- Mitchell & Krakauer (2022) The Debate Over Understanding in AI's Large Language Models
- Bender et al (2021) On the dangers of stochastic parrots
- Bender and Koller (2020) Climbing towards NLU
- Searle (1980) Minds Brains and Programs
- Baria and Cross (2021) The brain is a computer is a brain
- Hayles (2019) Can computers create meanings?
- Messeri & Crockett (2024) Artificial Intelligence and illusions of understanding in scientific research
- Grimm et al (2017) Explaining Understanding (esp Ch 1: "What is understanding?")
- Grimm (2021, SEP) Understanding
- Otsuka (2023, YouTube) What machine learning tells us about the mathematical structures of concepts
- Estrada (2024) Can a robot hand grasp?

Lesson 10: Algorithmic bias

Intro to AI Ethics:

- Bender (2022) Resisting dehumanization in the age of AI
- Asaro (2006) What should we want from a robot ethic?
- Mullaney et al (2021) Your Computer on Fire (Intro, Ch 1, 2, 6, 7, 8, 9)
- Crawford (2021) *The atlas of AI* (Intro, Ch 2, 3, 4)
- Whittaker (2021) The steep cost of capture

Algorithmic Bias

- Coded Bias (2020) documentary
- Buolamwini and Gebru (2018) Gender Shades
- Angwin et al (2016) Machine bias in sentencing
- Virginia Eubanks (2018) Automating Inequality (Intro, Ch 1, 2, 5)
- Benjamin (2019) Race after technology (Intro, Ch 1, 3)
- Noble (2018) Algorithms of Oppression Intro, Ch 1
- Cave & Dihal (2020). The whiteness of AI
- AI Now (2024): AI Nationalisms

Lesson 11: Autonomous weapons and vehicles

Estrada **Lecture Slides**
Autonomous weapons

- Roff and Moyes (2016) [Meaningful Human Control, Artificial Intelligence, and Autonomous Weapons](#)
- Amoroso & Tamburrini (2020) [AWS and Meaningful Human Control](#)
- Sharkey (2018) [AWS, Killer Robots, and Human Dignity](#)
- Asaro (2016) [Autonomous weapons](#)

Autonomous vehicles

- Benn Jordan (2024, Youtube) [Sorry, your car will never drive itself](#)
- Lin (2016) [Why Ethics Matters for Autonomous Cars](#)
- Koopman and Wagner (2017) [Autonomous Vehicle Safety](#)
- [MIT: Moral Machine](#) (2017 [Publication](#))
    - Jacques (2019) [Why The Moral Machine is a Monster](#)
- NHTSA Topic Overview: [AV Safety](#)

Lesson 12: Creativity and copyright

- Jaing et al (2023) [AI Art and its Impact on Artists](#)
- Natale and Hendrickson (2022) [The Lovelace effect: Perceptions of creativity in machines](#)
- Coeckelbergh and Gunkel (2023) [ChatGPT: deconstructing the debate and moving it forward](#)
- Gunkel (2021) [Computational Creativity](#)
- Ward (2020) [Victorian fictions of computational creativity](#)
- Samuelson (2023) [Generative AI meets copyright](#)
- Fenwick and Jurcys (2023) [Originality and the future of copyright in an age of generative AI](#)
- Franceschelli and Musolesi (2022) [Copyright in generative deep learning](#)

Lesson 13: Can or should robots have rights?

- [Universal Declaraction of Human Rights](#)
- [US Bill of Rights](#)
- Danaher (2017) [Should Robots have Rights? Four perspectives](#)
- Danaher (2020) [Welcoming robots into the moral circle: a defense of ethical behaviorism](#)
- Gunkel (2016) [Can and should robots have rights?](#)
- Gunkel (2018) [Robot Rights](#)
- Gunkel (2023) [Person, Thing, Robot](#)
- Bryson (2017) [Of by and for the people: the legal lacuna of synthetic persons](#)
- Birhane et al (2024) [Debunking robot rights metaphysically, ethically, and legally](#)
- Estrada (2017) [Robot rights: cheap, yo!](#)

Lesson 14: Ethical treatment of robots

- Darling (2015) ["Who's Johnny?" Anthropomorphic framing in human-robot interaction, integration, and policy](#)

- Darling (2016) Extending legal protection to social robots
- Bryson (2010) Robots should be slaves
- Salvini (2016) How safe are robots in urban environments? Bullying a service robot
- Estrada (2020) Human supremacy as posthuman risk
- Estrada (2017) Alignment, fair play, and the rights of service robots
- Estrada (2018) Sophia and her critics