



TEL AVIV אוניברסיטת תל אביב
UNIVERSITY תל אביב

Full Syllabus



Course Title:		Credits: 2
Introduction to Philosophy of AI		
Lecturer:		
Dr. Erez Firt		
Contact Details:		
erezfirt@gmail.com		
Semester		
Short Description:		
The course 'Introduction to Philosophy of AI' deals with philosophical questions related to the state of artificial intelligence today and in the near and more distant future. Its goal is to expose students to central concepts and ideas in the field of philosophy of artificial intelligence and machine ethics, including the difficulties we face when developing autonomous machines, as the importance of the decisions we delegate to them is increasing. The first part of the course will focus on building a conceptual and theoretical foundation - what is artificial intelligence, what is meant by narrow and general artificial intelligence, the different approaches to implementing artificial intelligence, what is involved in transferring decision-making to machines, the concept of moral agent, types of artificial moral agents in literature, and more. Building on this basis, we will move on to deal with topics related to the future of artificial intelligence, the significant problems facing us when we come to examine the development of future machines and artificial moral agents - the Control problem and the Value Alignment problem, the various attempts to solve these problems, and potential issues that will arise if and when we want to grant future machines moral status – safety, legal, and social problems.		
Course Outline		
1. Introduction: Course topics introduction, their importance, and a general introduction: What is artificial intelligence, brief history, different approaches (rule-based, statistical learning), narrow AI - where we are today, what can be achieved through artificial intelligence and what are its strengths, what cannot be achieved and why, weaknesses of existing methods (particularly in contexts of machine learning, ethical problems, lack of world understanding and common-sense), what is artificial general intelligence, what is necessary to achieve it. 2. Artificial Moral Agency: What is an agent? What is a moral agent? What are the necessary and sufficient conditions for moral agency, how to distinguish between different degrees of moral agency, what is an Artificial Moral Agent, the classification of different types of moral agents, the characteristics of each type, and the methods (mostly speculative ones) to create instances of these types. 3. Advance near-future AI: Ethical-social problems in contemporary artificial intelligence systems, the path to Human Level AI (HLAI), the next step - robust AI, why it is necessary and what is required to achieve it. What is understanding, what are the conditions for creating artificial understanding, the Intelligence Explosion argument, self-modifying systems, superintelligence/singularity, supporting and opposing views, opportunities and risks involved in developing Strong AI systems. 4. The Control Problem: What is the control problem, why this problem will worsen as artificial intelligence becomes more sophisticated and autonomous, attempts to solve the control problem (off switch, sandbox, and more) and examination of the likelihood of success for such solutions. 5. Value Alignment: What is value alignment, value alignment as a solution to the control problem, contemporary (degenerate) value alignment, problems on the path to value alignment.		

Full Syllabus

- Artificial General Intelligence (AGI): What is AGI, how and how not to define it (theoretically, practically), what basic capabilities are necessary and sufficient for AGI, what are the risks and opportunities, will we know how to identify it?
- Advanced topics: TBD.

Final grade components

Final writing assignment

Course schedule

This schedule is tentative and may change as the course progresses

Required course reading

Bostrom, N. (2003). "Ethical Issues in Advanced Artificial Intelligence". Available online, Nick Bostrom's site.

Himma, K. E. (2009). "Artificial agency, consciousness, and the criteria for moral agency : what properties must an artificial agent have to be a moral agent?". *Ethics and Information Technology*, 11: 19–29.

Moor, J. H. (2006). "The nature, importance, and difficulty of machine ethics". *Intelligent Systems, IEEE*, 21(4), 18–21.

Russell, S. (2017). "Provably Beneficial Artificial Intelligence". Available online, Stuart Russell's papers on Berkeley's edu site.

Optional course reading

Allen, C., Smit, I., & Wallach, W. (2005). "Artificial morality: Top-down, bottom-up, and hybrid approaches". *Ethics and Information Technology* 7:149 – 155.

Allen, C., Wallach, W. (2011) "Moral Machines: Contradiction in Terms, or Abdication of Human Responsibility?". In P. Lin, K. Abney, G. Bekey (Eds.) Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge: MIT Press, 55-68.

Armstrong, S., Sandberg, A. and Bostrom, N. (2012). "Thinking Inside the Box: Controlling and Using an Oracle AI". *Minds and Machines*. 22 (4): 299–324.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.



Full Syllabus



- Bostrom, N. (2012). "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents". *Minds and Machines*, 22(2): 71– 85 .
- Christiano, P. (2018). "Clarifying 'AI alignment'". Available online, ai-alignment's site.
- Everitt, T., Lea, G., & Hutter, M. (2018). AGI Safety Literature Review. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*. AAAI Press, 5441–5449.
- Firt, E. (2017). "Motivational defeaters of self-modifying AGIs". *Journal of Consciousness Studies*, Vol. 24/5-6: 150-169.
- Firt, E. (2020). "The Missing G", *AI&Society* 35: 995–1007.
- Firt, E. (2023). "Calibrating machine behavior: a challenge for AI alignment". *Ethics Inf Technol* 25, 42. <https://doi.org/10.1007/s10676-023-09716-8>.
- Goertzel, B. (2014). "Artificial General Intelligence: Concept, State of the Art, and Future Prospects". *Journal of Artificial General Intelligence* 5(1) 1-46.
- Good, I. J. (1965). "Speculations concerning the first ultraintelligent machine." In *Advances in Computers* 6, Franz L. Alt, and Morris Rubinof (eds.). Cambridge, MA: Academic Press.
- Haksar, V. (1998). "Moral agents". In *The Routledge Encyclopedia of Philosophy*. Taylor and Francis .
- Marcus, G. (2020). "The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence". <https://arxiv.org/abs/2002.06177>
- Marcus, G. (2022). "Deep Learning Is Hitting a Wall". <https://nautil.us/deep-learning-is-hitting-a-wall-14467/>, Accessed on 13-May-2022.
- Marcus, G. and Davis, E. (2020). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Vintage Books, NY.
- Müller, V. C. (2020). "Ethics of Artificial Intelligence and Robotics". In *Stanford Encyclopedia of Philosophy*..
- Russell, S. (2015). "Will they make us better people?" In J. Brockman (ed.), *What to think about machines that think?* (Edge Foundation), 359-362.
- Russell, S. (2019). *Human Compatible AI and the Problem of Control*. Penguin Random House LLC .
- Schlosser, M. (2019). "Agency". In *Stanford Encyclopedia of Philosophy* .
- Sezener, C. E. (2015). "Inferring human values for safe AGI design". In *Proceedings of the 8th International Conference on Artificial General Intelligence*, Berlin, Germany, 152-155. Springer International Publishing.
- Sullins, J. P. (2006). "When is a Robot a Moral Agent?" *International Review of Information Ethics*, 6(12): 23–29.
- Yampolskiy, R. (2012). "Leakproofing the Singularity Artificial Intelligence Confinement Problem". *Journal of Consciousness Studies* 19(1-2): 194–214.
- Yudkowsky, E. (2016). "AI Alignment: Why It's Hard, and Where to Start". Available online, Machine Intelligence Research Institute website.
- Yudkowsky, E. (2008). "Artificial Intelligence as a Positive and Negative Factor in Global Risk." In *Global Catastrophic Risks*, edited by Nick Bostrom and Milan M. Ćirković, 308-345. New York: Oxford University Press.

Academic Conduct

Plagiarism is taken extremely seriously. Any instance of academic misconduct which includes: submitting someone else's work as your own; failure to accurately cite sources; taking words from another source without using quotation marks; submission of work for which you have previously received credit; working in a group for individual assignments; using unauthorized materials in an exam and sharing your work with other students. In addition, [submitting a written](#)



TEL AVIV אוניברסיטת
UNIVERSITY תל אביב

Full Syllabus



assignment generated by AI in part or in whole without specific permission of the instructor is a form of plagiarism. Using any of the above will result in failure of the assignment and will likely lead to further disciplinary measures.

Attendance

Attendance is mandatory. The Program understands that there are occasions on which it is not possible for students to attend classes and/or unplanned events such as illness and outside commitments can prevent students from reaching campus. For this reason, *if absolutely necessary*, we allow students to miss up to four (4) sessions of each 4-credit course per semester and up to two (2) sessions for each 2-credit course. *These 2 or 4 allowable absences do not need to be authorized by the Liberal Arts office*, but students should inform their instructors as soon as they know they will need to be absent. Note that this allowance *includes* short-term illness, *even with presentation of a doctor's note*. Students who exceed these absences will face a grade penalty, and in extreme cases may not be able to complete the course/submit the final assignment or exam*. Note that individual instructors *are not authorized* to allow any additional absences.

Final assignment

Exam dates and submission deadlines are published on the Liberal Arts website. Courses with a final exam also have a second make-up exam. Late papers will be subject to a point deduction.^{1*} Students requiring an extension for a final paper must submit an Academic Committee Request Form to the Liberal Arts office in advance of the deadline.

Comments

^{1*}See Liberal Arts academic handbook for further details