**Philosophy 143: AI and the Future of Humanity**
Spring 2020, UNC Chapel Hill

TR 8:00am – 9:15am
Gardner 007

Instructor: Dominik Berger
Email: dominik@live.unc.edu
Office: Caldwell Hall 12B
Office Hours: Tuesdays 9.30am - 10.30am
Thursdays 11.00am - 12.00pm, or by appointment

**Course Description**
This course investigates philosophical issues arising from advanced forms of technology, in particular artificial intelligence and biological augmentation. We will consider questions about the dangers and benefits of AI, survival in non-biological ways, moral constraints on AI, the relationship between human and machine morality, and others.

**Readings**
You will need to purchase a copy of the following text:

*Superintelligence: Paths, Dangers, Strategies* by Nick Bostrom

All other readings will be made available on Sakai.

**Course Requirements**

Your grade will be calculated as follows:

Four Reading Reactions: 20%
Two Short Papers: 30% (each paper is worth 15%)
Forum Participation: 15%
One Midterm Exam: 15%
One Final Exam: 20%

**Reading Reactions:**
You will be required to submit **four** reading reactions over the course of the semester to readings that I will assign to you. The reading reactions should quickly, in two or three sentences (but no more), explain the main ideas behind a day's reading and raise two questions or objections about it. The purpose of this exercise is mainly for you to engage with an unfamiliar philosophical text and (1) try to figure out what's the main idea that the author is trying to convey (even if you might not understand all of the details), and (2) try to determine out which parts of the author's argument seem confusing or wrong to you. The reading reactions are **due at 8pm the day before** the class in which we will discuss the reading. I will distribute the reading reaction schedule after the first week of class.

**Short Papers:**
You will be required to write **two** short papers (around 3 pages) over the course of the semester. These papers are due on the dates outlined in the syllabus and will ask you to explain a philosophical view from the readings and offer a critical evaluation of an author's view. I will distribute prompts for these papers in advance of the due date.

**Forum Participation:**
I will assign you to one of 7 groups. Each group has their own discussion form in which to discuss the readings of a particular week. In this forum you should (i) discuss and answer the discussion questions I post for every class, (ii) raise additional questions about the readings that you have, and (iii) respond to each other's posts. I will give you a forum participation grade every week — so all your forum posts have to be written before Sunday of a particular week in order to receive credit.

**Midterm:**
There will be a take-home midterm covering the material for the first 6 or 7 chapters in Nick Bostrom's book that is due on February 6th. I will distribute the exam a week in advance.

**Final Exam:**
The final exam for this course will be a take-home exam that is due on Tuesday, April 28th at 8.00am. It will cover the material since the first midterm and consist of short answer questions that are designed to test your understanding of the main points discussed in the second part of the course. I will distribute the exam a week in advance.

**Note that the required writing for this course (including reading reactions, short papers and take-home midterm) will exceed 10 pages.**

**Due Dates and Late Policy**
I am usually happy to grant short extensions for the **short papers** and the take-home midterm as long as you send me an email and explain your situation. Late papers will be docked 1/3 of a letter grade for each day that they are late, unless you were experiencing a serious and genuinely unforeseen medical or personal emergency, and only when the emergency can be verified with the Dean of Students Office.
Reading reactions that are submitted *after* we discuss the relevant reading in class **won't** be accepted and will receive a grade of 0. If you are unable to submit a reading reaction on time due to a genuinely unforeseen medical or personal emergency that can be verified with the Dean of Students Office, I will assign you a different date for your reading response.
The final exam will take place on April 28th. Any request to take the Final Exam at a time other than the scheduled time must go through the Office of the Dean of Students (unless you have three or more exams scheduled within a 24 hour period).

**Note: Please write your PID instead of your name on your papers and exams to allow for blind grading.**

**Accommodations**
If you require reasonable accommodations for a documented disability, you must register with ARS (https://accessibility.unc.edu/). Once I receive ARS's recommendations, I will be happy to work with you to implement them as appropriate.

**Laptops**
In my experience laptops are distracting both to you and the people around you, so laptops, phones, and other electronics are not permitted for use in the classroom. If there is a special reason for which you have to use electronics and you let me know (either in person or through ARS) you are of course permitted to use them.

**Outside Sources**
Please do not refer to any academic sources other than the assigned readings in your papers. The one exception is the Stanford Encyclopedia of Philosophy (available at http://plato.stanford.edu/), which is very useful for general background reading on philosophical terms and topics.

**Plagiarism**

The UNC Instrument of Student Governance defines plagiarism as "deliberate or reckless representation of another's words, thoughts, or ideas as one's own without attribution in connection with submission of academic work, whether graded or otherwise." You are expected to abide by UNC's Honor Code, and refrain from any kind of academic dishonesty, including cheating and plagiarism. Just as you are bound by the Honor Code not to plagiarize, I am bound by it to report suspected cases of academic dishonesty of any kind to the Honor Court.

In your papers, you may use whichever standard citation convention that you'd like. But any words that you borrow from any external source must appear in quotation marks, and you must provide some sort of internal citation indicating where those words came from. It is also a form of plagiarism to closely paraphrase text from an external source without proper citation, changing a few of the words but imitating the structure of the external source.

In addition, please bear in mind that plagiarism can be committed non-deliberately; if you are reckless in your use of other people's words or ideas, then you have committed plagiarism even if you didn't mean to do so. If you have any questions at all about proper citation of other people's words or ideas in the course, please don't hesitate to come talk to me about them. You are responsible for knowing what exactly counts as plagiarism and to not commit it in your papers.

(Note: This Syllabus is still provisional and subject to change.)

## Course Schedule

### Week 1 — Introduction

Thursday, January 9th — Introduction
No reading.

### Week 2 — Unit 1: Superintelligence

Tuesday, January 14th — A short history of AI developments
Read Bostrom's *Superintelligence*, Chapter 1

Thursday, January 16th — Different ways in which Superintelligence could be achieved
Read Bostrom's *Superintelligence*, Chapter 2

### Week 3 — Unit 1: Superintelligence

Tuesday, January 21st — Different forms Superintelligence could take
Read Bostrom's *Superintelligence*, Chapter 3

Thursday, January 23rd — What might happen once we create Superintelligence
Read Bostrom's *Superintelligence*, Chapter 4

**Week 4 — Unit 1: Superintelligence**

Tuesday, January 28th — What might happen once we create Superintelligence, Part 2
Read Bostrom's *Superintelligence*, Chapter 5

Thursday, January 30th — What might happen once we create Superintelligence, Part 3
Read Bostrom's *Superintelligence*, Chapter 6

**Week 5 — Unit 1: Superintelligence**

Tuesday, February 4th — What kinds of things might a superintelligent agent want?
Read Bostrom's *Superintelligence*, Chapter 7

Thursday, February 6th — A superintelligent agent might act ethically
Read Peterson's "Superintelligence as superethical"
**Midterm due.**

**Week 6 — Unit 1: Superintelligence**

Tuesday, February 11th — The Control Problem
Read Bostrom's *Superintelligence*, Chapter 9

Thursday, February 13th — The Control Problem, Part 2
Read Bostrom's *Superintelligence*, Chapter 10

**Week 7 — Unit 1: Superintelligence**

Tuesday, February 18th — The value-loading problem
Read Bostrom's *Superintelligence*, Chapter 12

Thursday, February 20th — Beneficial AI
Read Russell's *Human Compatible*, Chapters 7 and 8

**Week 8 - Unit 2: Consciousness**

Tuesday, February 25th — Identity Theory: The mind is the brain
Read Place's "Is consciousness a brain process?"

Thursday, February 27th — Functionalism
Read Putnam's "The nature of mental states"
**Paper 1 on Superintelligence due.**

**Week 9 — Unit 2: Consciousness**

Tuesday, March 3rd — Dualism
Read Chalmer's "Facing up to the problem of consciousness"

Thursday, March 5th — What would alien minds be like?
Read Schneider's "Alien Minds"

**Week 10 — Spring Break**

Tuesday, March 10th — Spring Break

Thursday, March 12th — Spring Break

**Week 11 — Spring Break**

Tuesday, March 17th — Spring Break

Thursday, March 19th — Spring Break

**Week 12 — Unit 3: Personal Identity and Uploading**

Tuesday, March 24th — Personal Identity Primer
Read Sider's "Personal Identity"

Thursday, March 26th — Personal identity isn't important for survival
Read Parfit's "Personal Identity"

**Week 13— Unit 3: Personal Identity and Uploading/Unit 4: Moral Status**

Tuesday, March 31st —Would we survive uploading our minds?
Read Chalmer's "Mind Uploading — A philosophical analysis"

Thursday, April 2nd — What grounds one's moral status
Read SEP's "The grounds of moral status", Section 5

**Week 14 — Unit 4: Moral Status**

Tuesday, April 7th —Should AI have rights?
Read Schwitzgebel and Garza's "A defense of the rights of artificial intelligence"

Thursday, April 9th — Is consciousness important for moral status?
Read Levy's "The value of consciousness"
**Paper 2 on Consciousness/Personal Identity due.**

**Week 15 — Unit 5: Living with enhanced beings**

Tuesday, April 14th —What would it mean for us if we lived together with enhanced humans?
Read Douglas's "Human enhancement and supra-personal moral status"

Thursday, April 16th — What would it mean for us to be the last moral humans?
Read Rini's "The last mortals"

**Week 16 — Unit 5: Living with enhanced beings**

Tuesday, April 21st — Would it matter if we lived in a simulation?
Read Chalmers's "The virtual and the real"

Thursday, April 23rd - What matters for a meaningful life?
Read Wolf's "Happiness and meaning"

**Week 17 - Final exam**

Tuesday, April 28th — Final Exam at 8.00am