Course Title: Ethics for Artificially Intelligent Robots: A Practical Philosophy for Our Technological Future  
Course Code: PHI 114  
Instructor: Forrest Hartman

Course Summary: The rise of intelligent robots, an unprecedented social phenomenon, is an eventuality that will not be postponed and calls for new ways of thinking about how what is so alien can be integrated successfully and smoothly into society. A critical marker of the social is the capacity for ethical behavior, itself an indicator of intelligence. This course specifically questions the very possibility of robot ethics, even as a practical extension of what we usually think of as a uniquely human attribute. What should we be looking for in and expect from robot ethics? If we are to trust autonomous machines in our midst--machines that can act on their own--we must be confident that they can and will act ethically and will do no harm. This course approaches robot ethics as an interdisciplinary investigation that calls on every facet of our common cultural enterprise to prepare for what will be the greatest transformation of human society in history.

Grade Options and Requirements:
- No Grade Requested (NGR)
  - This is the default option. No work will be required; no credit shall be received; no proof of attendance can be provided.
- Credit/No Credit (CR/NC)
  - Score will be determined by student attendance and participation. Whoever chooses this option needs to sign in on an attendance sheet each week.
- Letter Grade (A, B, C, D, No Pass)
  - Written work, a 5 to 10 page paper, to be arranged with the instructor and to reflect the particular interest of the student, will determine the grade.

*Please Note: If you require proof that you completed a Continuing Studies course for any reason (for example, employer reimbursement), you must choose either the Letter Grade or Credit/No Credit option. Courses taken for NGR will not appear on official transcripts or grade reports.

Tentative Weekly Outline:
Most of the readings are from Machine Ethics (edited by Michael Anderson and Susan Leigh Anderson), the textbook for the course, which will be available on 31 March. The first two assignments will be available on Canvas, after which you should be able to obtain the book. Where indicated other readings can be found in the two recommended books (Superintelligence and Life 3.0).
I

Preparatory Topics:
Preparatory Considerations,
Prerequisite to Approaching the Bizarre Topic of Machine Ethics

Week 1: 4 April
What can we expect from robot ethics as a disciplined field of inquiry and what it could even mean to introduce (upload) anything like ethics (a uniquely human concern) into a machine.

1. “General Introduction” and “Introduction” (pp. 1-12)
3. “Machine Metaethics,” Susan Leigh Anderson (pp. 21-27)
4. “Ethics for Machines,” J. Storrs Hall (pp. 28-44)

Week 2: 11 April
Why an inquiry into robot ethics now is so urgent that it can no longer be postponed in light of the acceleration of the development of artificial general intelligence and why, concomitant with research in general artificial intelligence, we also need to focus our attention on this newly-emerging discipline of robot ethics.

1. “Introduction” (pp. 47-50)
3. “Authenticity in the Age of Digital Companions,” Sherry Turkle (pp. 62-76)

Week 3: 18 April
Why we must question the very possibility of a robot ethics as an applied ethics that is extended beyond the human to an artificial device, the very status of which in the general scheme of things is questionable and why we could never trust an autonomous machine that lacked an ethical sensibility.


A.

Will the very presence of intelligent robots demand that we totally revise our narrow, anthropocentric way of thinking about ethics, or does “ethics” become an arbitrary term if we try to extend it to the artificial?
1. “Introduction,” (pp. 79-87)
2. “What Matters to a Machine?” Drew McDermott (pp. 88-114)

Week 4: 25 April


B.

If not fully ethically equivalent to human beings, could robots nonetheless still have an ethical function, for example, as advisors, as “moral entities,” even if they are not
fully capable autonomous, co-responsible moral agents? Could they help coordinate and share ethical responsibilities or even confirm ethical decisions, perhaps superseding the prejudices and biases to which human ethical decisions are prone?

1. “On Computable Morality: An Examination of Machines as Moral Advisors,” Blay Whitby (pp. 138-150)
2. "When is a Robot a Moral Agent?" John P. Sullins (pp. 151-161)
3. “Philosophical Concerns with Machine Ethics,” Susan Leigh Anderson (pp. 162-167)

II

Conjuring Up the Ethical Robot:
Envisioning Our Future Co-Existence with Social Robots

Week 5: 2 May


A. Science Fiction as Thought Experiments

The “brute-force” ethics of rule-following as prescriptive rules or imperatives, the most intuitive approach for computational thinking: the deontological approach as investigated through thought experiments, in particular, those dramatized in the science fiction of Isaac Asimov.

1. “Introduction” (pp.231-243)
2. “Towards the Ethical Robot,” James Gips (pp. 244-253)
3. “Asimov’s Laws of Robotics Implications for Information Technology, “ Roger Clarke (pp. 254-284)
5. The philosophical critique of rule-following in Ludwig Wittgenstein’s Philosophical Investigations. (excerpts)

Week 6: 9 May


B. Modeling and the Technical Approaches of AI Research

The computational, reductionist approach and the technical resources of computational and neural network artificial intelligence research: Modelling the ethical robot as the quintessential computational device and the limits of neural networks for “learning” ethics.

1. “Computational Neural Modeling and the Philosophy of Ethics Reflections on the Particularism-Generalism Debate,” Marcello Guarini (pp. 316-334)
2. “Architectures and Ethics for Robots Constraint Satisfaction as a Unitary Design Framework,” Alan K. Mackworth (pp. 335-360)
**Week 7: 16 May**


**C. The Contributions of Traditional Philosophy**

The extent to which traditional, abstract philosophical approaches to ethics are applicable or relevant to a viable robot ethics, with particular reference again to the intuitive allure of utilitarianism and Kant and a review of earlier philosophical concepts.

1. “There is no ‘I’ in ‘robot’: Robots and Utilitarianism,” Christopher Grau (pp.451-463)
2. “Prospects for a Kantian Machine,” Thomas M. Powers (pp.464-475)

**III**

**Charting a Course Ahead:**

_Where to Go from Here as a Matter of How We Want to Be Situated in Our Technological Future with Superintelligent Robots_

**Week 8: 23 May**

*Superintelligence*

Unless ethics is reduced to the simplistic problem of downloading rules (as if merely a program installed for data manipulation), could other means be found for instilling ethics into a machine, and how would our chosen method affect the fundamental values we chose to be inculcated?

1. “Acquiring Values,” Chapter 12 (pp. 226-255)
2. “Choosing the Criteria for Choosing,” Chapter 13 (pp. 256-279)

**Week 9: 30 May**

*Life 3.0: Being Human in the Age of Artificial Intelligence*

How, at a still more profound level, ethics is necessarily founded on commonly-held values that reflect the way we want to live and how that primordial, even ineffable decision will determine the goals which _prospectively_ will _pre-_scribe the ethics that we instantiate in our ultimate technological achievement.

1. “Aftermath: The Next 10,000 years,” Chapter 5 (pp. 161-202)
2. “Goals,” Chapter 7 (pp. 249-280)

**Week 10: 6 June**

Prospects for the possibility and _necessity_ of a compelling and viable robot ethics: Charting the course ahead.

1. _Superintelligence:_ “Crunch Time,” Chapter 15 (pp. 314-320) and “Afterward” (pp. 321-324)

Can the discipline of robot ethics lead to a better understanding of human ethics, or will it call for a total reconception of the ethical as such? Similar to the way that computer scientists had once believed that we could better understand human reasoning on the analogy of
mind is to brain as program is to computer, will we once again try to delimit human ethics in terms of what is fitting for robots?

i. “Introduction” (pp. 495-498)

ii. “What can AI do for Ethics?” Helen Seville and Deora G. Field (pp. 499-511)

iii. “Ethics for Self-Improving Machines,” J. Storrs Hall (pp. 512-523)

iv. “How Machines Might Help Us Achieve Breakthroughs in Ethical Theory and Inspire Us to Behave Better,” Susan Leigh Anderson (pp. 524-530)

v. “Homo Sapiens 2.0: Building the Better Robots of Our Nature,” Eric Dietrich (pp. 531-538)