Philosophical Questions Through a Cinematic Lens I:
AI in Science Fiction

(ABRIDGED) COURSE OUTLINE

Detailed Course Description.

This course proceeds from the view that cinema provides a rich medium for raising and addressing philosophical problems. In the first of what will be a set of courses on different philosophical themes, we will centre on artificial intelligence in science fiction films.

Almost 100 years before the phrase 'artificial intelligence' entered the scientific lexicon, and nearly 75 years before the first programmable computer was invented, there was already literary reference to machines with consciousness (Samuel Butler, 1863, 'Darwin and the Machines'). In the last 100 years, there have been more than 100 films in which artificial intelligence plays some role—the earliest being *Metropolis* (1927, dir. Fritz Lang). Many of these films have addressed philosophical problems that have become truly pressing in the last decade, with the advent of deep learning bolstered by 'big data'; remarkably, many of the concepts and problems raised pre-figure any real-world implementation. Thus, in this course, we will use fictional cinema as a gateway into thinking philosophically about real ethical, social, and conceptual problems that arise from emerging technologies in the present day. Each week there will be (typically) one required film and (typically) one required reading that has been chosen to accompany some (and by no means all) of the philosophical themes of that week's film.

In the first part of the course, we will be looking at broad, conceptual questions that arise in light of the possibility of artificial intelligence. For example,

- What does it mean to be human? (*Alphaville*, 1964)
- What does the Turing test tell us? How can we know whether an AI is an agent? (*Ex Machina*, 2015)
- Can an AI be a moral patient, worthy of moral consideration? (*Blade Runner*, 1982)
- What if we cannot control theAI systems we create? (*2001: A Space Odyssey*,)
- What if the objectives of the systems we create are mis-aligned with our own values and goals? (*The Matrix*, 1999)

In the second part of the course, we will direct our focus to specific ethical and social consequences that might arise from technology. Themes will include,

- Cognitive Enhancement (*Eternal Sunshine of the Spotless Mind,* 2004 and *Black Mirror* S1E3, 2011);
- Virtual Reality (*eXistenZ*, 1999);
- How AI systems might affect policing, and whether this is fair or ethical (*Minority Report*, 2002);
- How technology may increase inequity in society (*Elysium*, 2013);
- How human social relations may change in light of artificial agents (*Her*, 2013); and,
- How (or whether) AI may be used for good (*WALL-E*, 2008).

Content Warning.

Due to the nature of this course, some of the content may be disturbing. This will typically include moderate or significant graphic violence, mild to moderate language, and mild to moderate sex or nudity.

Most of the films are rated 14A in Canada (13+ Québec) or lower, with the one exception being eXistenZ, which is rated R in Canada (13+ in Québec). With the exception of *Ex Machina*, there are no depictions of self-harm, graphic misogyny, or sexual assault.

Learning outcomes.

This course aims at providing you with a clear understanding of core philosophical concepts and issues in the philosophy of artificial intelligence, as seen through the lens of science fiction cinema. This course will enable you to develop and sharpen your interpretative and analytical skills, using the cinematic medium. You will learn to articulate your insights and thoughts on pressing philosophical problems surrounding the advent of artificial intelligence in clear and rigorous terms. You will also learn to discuss philosophical issues and concepts in relation to the film medium.

GRADING DETAILS

Participation. (20 marks total) Timely arrival, attendance, and engagement in lecture and during screenings/discussion will count toward the participation mark.

Short Paper 'Reflections'. (8 x 10 marks total) Over the course of the semester, we will be watching 12 feature films. Out of these twelve weeks, you are required to complete 8 short papers (300-600 words) which address the concepts explored in the screening for that week.

Bonus Marks. There will be two opportunities for obtaining bonus marks in this course. One bonus mark will be awarded for a perfect score on a quiz on the content of this syllabus (completed on the first day of class). A second bonus mark will be awarded to everyone registered just in case a quorum (at least 2/3) of students completes the year-end course evaluations. More details for each of this are given below.

Overview. The final grade for this course is broken down as follows:

Participation	20 %
Weekly Short Paper 'Reflections'	80 %
Bonus Marks	2 %
TOTAL	**100 %**

Topic Overview

Part I: Theoretical Problems
Week 1: Course Introduction
Week 2: Humanity, Intelligence, and Agency
Week 3: Artificial Moral Agents
Week 4: Artificial Moral Patients
Week 5: The Control Problem
Week 6: The Alignment Problem
Week 7: Malevolent AI
Week 8: *Reading Break* (No Class)

Part II: Applied Problems
Week 9: Cognitive Enhancement
Week 10: Virtual Reality
Week 11: Policing
Week 12: Inequity
Week 13: Sex, Love, and Social Relations
Week 14: AI for Good
Week 15: *Exam Period I*
Week 16: *Exam Period II*

Film Overview

Part I: Theoretical Problems
Week 1: *The Intelligence Explosion*
Week 2: *Alphaville*
Week 3: *Ex Machina*
Week 4: *Blade Runner*
Week 5: *2001: A Space Odyssey*
Week 6: *The Matrix*
Week 7: *Terminator 2: Judgment Day*

Part II: Applied Problems
Week 9: *Eternal Sunshine of the Spotless Mind*
Black Mirror (S1, E3)
Week 10: *eXistenZ*
Week 11: *Elysium*
Week 12: *Minority Report*
Week 13: *Her*
Week 14: *WALL-E*

Detailed Course Schedule

Note that there is typically one feature film per week, and there is typically one required reading per week. The optional readings are listed in the detailed schedule. You are ***not*** expected to read this supplementary material. However, the content of the supplementary material *may* be referenced during the lecture. So, complete references are provided in case you would like more detail, context, or clarification. It is permissible, but certainly *not obligatory*, to utilise the supplementary readings for your short papers.

Week 1	**Course Introduction**
Reading	Syllabus
Screening (In Class)	*The Intelligence Explosion* (Dan Susman, 2017, 5 min.)
	PART I: Theoretical Problems

Week 2 **Humanity, Intelligence, and Agency**

Reading
Alan M. Turing (1950) ‘Computing Machinery and Intelligence’ *Mind*. LIX(236): 433-460.

Screening
Alphaville (Jean-Luc Godard, 1965, 99 min.)

——————— ———

Optional
Jones, Andrew. 2017. 'Art and logic: Godard’s Alphaville as philosophy’, *Studies in French Cinema*. 17(2): 165-181.

Markus Schlosser. 2019. ‘Agency', in Edward N. Zalta (ed.) *Stanford Encyclopedia of Philosophy*.

Shaun Gallagher. 2007. ‘The Natural Philosophy of Agency’, *Philosophy Compass*. 2(2): 347-357.

Tim Bayne. 2008. 'The Phenomenology of Agency’, *Philosophy Compass* 3(1): 182-202.

Suvradip Maitra 2020. ‘Artificial Intelligence and Indigenous Perspectives: Protecting and Empowering Intelligent Human Beings’, in Annette Markham, Julia Powles, Toby Walsh, Anne L. Washington (eds.) *AIES ’20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery: 320-326.

Week 3 **Artificial (Moral) Agents**

Reading
John R. Searle. 1980. Minds, brains, and programs. *The Behavioral and Brain Science*. 3: 417-457.

Screening
Ex Machina (Alex Garland, 2015, 108 min.)

——————— ———

Optional
B. Jack Copeland. 2000. ‘The Turing Test’, *Minds and Machines* 10: 519-539.

Graham Oppy and David Dowe. 2020. ‘The Turing Test’ in Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/turing-test/

David Cole. 2020. ‘The Chinese Room Argument’ in Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/chinese-room/

Frank Jackson. 1982, ‘Epiphenomenal Qualia’, *Philosophical Quarterly* 32: 127-136.

Frank Jackson. 1986, ‘What Mary Didn’t Know’, *Journal of Philosophy*, 83: 291-295

Martine Nida-Rümelin and Donnchadh O Conaill. 2019. ‘Qualia: The Knowledge Argument’ in Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/qualia-knowledge/

Jason David Grinnell. 2020. ‘Ex Machina as Philosophy: Mendacia Ex Machina (Lies from a Machine)’, in D. Johnson (ed.) *Palgrave Handbook of Popular Culture as Philosophy*. Palgrave Macmillan, Cham. 1-18.

Week 4 **Artificial Moral Patients**

Reading S. Matthew Liao. 2020. The Moral Status and Rights of Artificial Intelligence' in S. Matthew Liao (ed.) *Ethics of Artificial Intelligence*. Oxford: Oxford University Press. 480-503.

Screening *Blade Runner* (Ridley Scott, 1982, 117 min.) [n.b. 'Final Cut' version from 2007]

Optional Timothy Shanahan. 2020. '*Blade Runner* as Philosophy: The Replicants Are Us (Almost)', in D. Johnson (ed.) *Palgrave Handbook of Popular Culture as Philosophy*. Palgrave Macmillan, Cham. 1-21.

Agnieszka Jaworska and Julie Tannenbaum. 2018. 'The Grounds of Moral Status' in Edward N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/grounds-moral-status/

Week 5 **The Control Problem**

Reading: Nick Bostrom. 2014. 'The Control Problem', Ch. 9 in *Superintelligence: Paths, Dangers Strategies*. Oxford: Oxford University Press. 155-176.

Screening: *2001: A Space Odyssey* (Stanley Kubrick, 1966, 164 min.)

Week 6 **The Alignment Problem**

Reading: Iason Gabriel. 2020. 'Artificial Intelligence, Values, and Alignment'. *Minds & Machines*. 30: 411–437.

Screening: *The Matrix* (Lana Wachowski, Lilly Wachowski, 1999, 136 min.)

Optional Thomas Arnold, Daniel Kasenberg, and Matthias Scheutz. 2017. 'Value Alignment or Misalignment — What Will Keep Systems Accountable?' In *3rd International Workshop on AI, Ethics, and Society*. AAAI Workshops. 1-8.

Osonde A. Osoba, Benjamin Boudreaux, and Douglas Yeung. 2020. 'Steps Towards Value-Aligned Systems. In Annette Markham, Julia Prowles, Toby Walsh, and Anne L. Washington (eds.) *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery: 332-336.

Norbert Weiner. 1960. 'Some Moral and Technical Consequences of Automation', Science 131(3410): 1355-1358.

Week 7 **Superintelligence and Malevolent AI**

Reading Stuart Russell. 2019. 'Intelligence in Humans and Machines', Ch. 2 in *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking. 13-61.

Screening *Terminator 2: Judgment Day* (1991, James Cameron, 137 min.)

Optional Nick Bostrom. 2014. 'Afterword', in *Superintelligence: Paths, Dangers Strategies.* Oxford: Oxford University Press. 321-324.

Week 8 Reading Break (*No Class*)

PART II: Applied Problems

Week 9 **Cognitive Enhancement**

Reading Richard Heersmink and J. Adam Carter. 2020. The philosophy of memory technologies: Metaphysics, knowledge, and values' *Memory Studies*. 13(4): 416-433.

Screening 'The Entire History of You' *Black Mirror* S1 E3 (Brian Welsh, 2011, 44 min.)

Eternal Sunshine of the Spotless Mind (Michel Gondry, 2004, 108 min.)

——————— ———

Optional José Hernéndez-Orallo and Karina Vold. 2019. 'AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI'. *Proceedings of the the 2019 AAAI/ACM Conference.*

Week 10 **Virtual Reality**

Reading Thomas K. Metzinger. 2018. "Why is Virtual Reality Interesting for Philosophers?" *Frintiers in Robotics and AI.*

Screening *eXistenZ* (David Cronenberg, 1999, 97 min.)

——————— ———

Optional Duncan Pritchard. 2012. 'eXistenZial Angst', in Simon Riches (ed.) *The Philosophy of David Cronenberg*. Lexington: University Press of Kentucky. 69-76.

Graham Stevens. 2012. The Fiction of Truth in Fiction Some Reflections on Semantics and eXistenZ', in Simon Riches (ed.) *The Philosophy of David Cronenberg*. Lexington: University Press of Kentucky. 143-154.

Week 11 **Inequity**

Reading Peter Singer. 1972. 'Famine, Affluence, and Morality'. *Philosophy and Public Affairs*, 1: 229-243.

David Rotman. 2014. 'Technology and Inequality' *MIT Technology Review*.

Christoph Lutz. 2019. 'Digital Inequalities in the Age of Artificial Intelligence and Big Data', *Human Behavior and Emerging Technologies* 1(2): 141-148.

Screening *Elysium* (Neill Blomkamp, 2013, 109 min.)

Week 12	**Policing**
Reading	Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. ‘Machine Bias’ *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
	Brian Christian. 2020. ‘Fairness’, Ch. 2 in *The Alignment Problem*. W. W. Norton & Company. 51-81.
Screening	*Minority Report* (Steven Spielberg, 2002, 146 min.)
———————	———————————————————————————————
Optional	Mahdi Hashemi and Margaret Hall. 2020. “RETRACTED ARTICLE: Criminal tendency detection from facial images and the gender bias effect”*Journal of Big Data* 7(2): 1-16.
	Kevin W. Bowyer, Michael C. King, Walter J. Scheirer, and Kushal Vangara. ‘The “Criminality From Face” Illusion’ *IEEE Transactions on Technology and Society*. 1(4): 175-183
Week 13	**Sex, Love, and Social Relations**
Reading	Kate Devlin. ‘The Ethics of the Artificial Lover’ in Matthew S. Liao (ed.) Ethics of Artificial Intelligence. Oxford: Oxford University Press. 271-290.
Screening	*Her* (Spike Jonze, 2013, 126 min.)
———————	———————————————————————————————
Optional	John P. Sullivan. 2012. ‘Robots, Love, and Sex: The Ethics of Building a Love Machine”, *IEEE Transactions on Affective Computing* 3(4): 398-409.
Week 14	**AI For Good**
Reading	Ben Green. 2019. ‘“Good” isn’t Good Enough’, in *Proceedings of the AI for Social Good Workshop at NeurIPS*.
	Mariarosaria Taddeo and Luciano Floridi. 2018. ‘How AI can be a force for good’. Science 361(6404): 751-752.
Screening	*WALL-E* (Andrew Stanton, 2008, 103 min.)
Week 15	**Exam Period I**
	(*No Class*)
Week 16	**Exam Period II**
	(*No Class*)