‘Artificial Intelligence and the Value Alignment Problem’

COURSE OUTLINE

1. Course Information

Course Times + Location	Exam Times + Location
TBA	N/A
TBA	N/A
Contact Information	Office Hours
Instructor: Dr Travis LaCroix (he/him)	TBA
Email:	*TBA*

1.1 Contact policies. My email policy is to respond to any enquiries within two work-days of receipt. If I have not responded to your email within this time frame, you are entitled to (and should) send a follow-up email. Please put the course code (‘PHIL xxxx’) in the subject-line of your email. For scheduled office hours, you are welcome to drop-in without letting me know in advance. If the set times do not work for you, I am also available by appointment, either in-person or via Zoom. Please send an email to set up a time.

1.2 Course Delivery Information. In light of the global pandemic, this course will consist of synchronous seminar meetings conducted via Zoom. Links to the meeting will be posted on the course webpage.

> *Note:* If you are auditing this course, or on a waitlist to register, and do not have access to the course web-page, please email me.

1.3. Detailed course description. Artificial intelligence research is progressing quickly, and along with it the capacities of AI systems. As these systems become more sophisticated and more deeply embedded in society, it will becomes increasingly essential to ensure that we are able to maintain control of these systems, and that the decisions and actions they take are aligned with the values of humanity writ large. These are known, in the field of machine ethics, as the *control problem* and the *value alignment problem*, respectively.

In the first part of the course, we will examine the concepts of control and value alignment to see how they are connected and what practical, ethical questions arise when trying to solve these problems. In the second part of the course, we will focus on the normative component of value alignment, which asks what values or principles (if any) we ought to encode in artificial agents—namely, how to achieve moral agency in an artificial system. We will discuss the three main approaches to artificial moral agency, which correspond roughly to the three main normative ethical theories in western philosophy—utilitarianism, deontology, and virtue ethics. In the final part of the course, we will examine the social, ethical, and philosophical consequences that might arise (indeed, have arisen) from misaligned AI systems.

1.4. Course Objectives and Learning outcomes.

- Deep knowledge of contemporary ethical problems surrounding value-aligned artificial intelligence (with respect to both conceptualisation and implementation).
- Ability to identify and articulate questions for discussion and investigation.
- Ability to critically digest, interpret, and analyze complex, multi-disciplinary sources.
- Ability to write a convincing argument that takes adequate account of alternative positions.
- Ability to engage in constructive, respectful, oral, and written discussion.
- Ability to use feedback about one's work to improve one's arguments and writings.
- Development of practical skills in professional philosophy.

1.5 Prerequisites. Determined by course coding.

1.6 Co-requisites. None.

1.7 Anti-requisites. None.

1.9 Required course materials. All of the required readings for this course will be made available online on the course webpage. *Note*: if you are auditing or on the waitlist and do not have access to the course web-page, please email me.

1.10 Territorial Acknowledgement. [Include territorial acknowledgment]

2. Course Assessment And Evaluation

The assignments for this course differ depending on whether you are registered in the graduate or undergraduate section. If a student registered in the undergraduate section is so inclined, it is possible for them to be graded according to the graduate scheme. In this case, they should state these intentions via email before the end of Week 3. Details for each of the sections are given below.

2.1 Grading (Undergraduate only)

Participation. (10 marks total) Timely arrival, attendance, and engagement in class will count toward the participation mark. It is expected that your contributions will be respectful and constructive. See ground rules below.

Weekly Reading Responses. (15 marks total) Students are required to write and submit a question on one of the assigned readings for the week by 1 pm the day before our meeting. Specific instructions will be available on the course webpage. Weekly questions and their explanations/context should not generally be longer than 250 words. Over the 12 weeks of the term for which there are readings assigned, you must submit **ten** (10) weekly questions; which weeks you submit is up to you. Submission of more than ten weekly reading responses will be considered for the participation grade.

Short Essays. (3 x 25 marks) The default requirement for those enrolled in the undergraduate section is three short papers (750-1250 words). These papers should isolate one localized philosophical, conceptual, or methodological point within the readings and offer some analysis and/or critique. The thesis and its defense need not be earth-shattering in any way; this is really just an exercise in finding a topic of the right size and crafting a thesis and defense to match. I encourage writers to email me their topic and thesis, or a draft introductory paragraph, for discussion well before the due date.

Bonus Marks. (2 marks total) There will be two opportunities for obtaining bonus marks in this course. Details for each of these are given below.

> *Syllabus Quiz***.** (1 bonus mark) On the first day of class, we will spend some time going through this syllabus in detail. There will be a 'quiz' (multiple choice) on the content of this syllabus, to be completed in groups and handed in individually. If (and only if) you receive a perfect score on the syllabus quiz, a bonus mark will be added to your final grade for the course. For example, if your final grade at the end on the semester is 89/100, and you received a perfect score on the syllabus quiz, you will receive a final grade of (89 + 1 =) 90/100. The syllabus quiz must be handed in at the end of the first class in order to be eligible for a bonus mark. More details will be given on the first day.
>
> *Course Evaluations Game***.** (1 bonus mark) If a 2/3 majority of students fill out the year-end evaluation, then everyone will receive one bonus mark for the course. Note that this bonus assignment has a structure typical of a prisoner's dilemma: If most students cooperate (fill out the evaluation), then it is in your individual interest to not (because you can get a bonus mark without expending additional effort in filling out the evaluation). Further, if most students defect (fail to fill out the evaluation), it is again in your best interest to defect (otherwise, you would have expended additional effort for nothing). This is a dilemma because it will always be in your own best interest to defect; however, it is in everyone's best interest to cooperate. Do with this information what you will.

2.2 Grading Overview (Undergraduate only). The grading for the undergraduate section of this course is broken down as follows:

Participation	10 %
Weekly Reading Responses	15 %
Short Paper 1 (1000 words)	25 %
Short Paper 2 (1000 words)	25 %
Short Paper 3 (1500 words)	25 %
Bonus Marks	2 %
TOTAL	**100 %**

2.3 Grading (Graduate section only)

Participation. (0 marks total) You will not be formally graded for participation, but registration in the graduate section of this course come with an expectation that you will be an active contributor to the discussion in the weekly meetings. It is also expected that your contributions be respectful and constructive. See ground rules below.

Weekly Reading Responses. (10 marks total) Students are required to write and submit a question on one of the assigned readings for the week by 1 pm the day before our meeting. Specific instructions will be available on the course webpage. Weekly questions and their explanations/context should not generally be longer than 250 words. Over the 12 weeks of the term for which there are readings assigned, you must submit **six** (6) questions on separate readings; which weeks you submit is up to you; however, your weekly reading responses may not be on the readings assigned for (1) the referee report assignments, or (2) the conference commentaries.

'Referee Reports'. (20 marks total) By week 2, you will be *assigned* responsibility for three (3) of the required readings, from which to choose two (2) to submit a 'referee report', as if you were revising the assigned paper for publication. That is, of the three papers assigned to you, you can choose two to provide a report for. You will be given an opportunity, in week 1 to submit your preferences for these assignments, but the assignments cannot be switched once they are made. These reports may be submitted at any time throughout the semester, up to the day before the meeting in which that reading is discussed. More details, as well as some resources to help you write a constructive report, will be provided on the course webpage.

'Conference Commentaries'. (20 marks total) By week 2, you will be *assigned* **three** (3) of the required readings from which you are to provide two 'commentary' presentations. These commentaries will be presented in class on the day for which that reading is assigned. As with the referee report assignment, you will be given an opportunity, in week 1 to submit your preferences. This assignment should be treated as though you were providing a commentary on a paper at a conference. As such, the presentation should be no more than 10 min. More details, as well as some resources to help you write and present a constructive commentary, will be provided on the course webpage.

Term Paper. (50 marks total) The term papers will be broken up into several components, including a proposal (5 marks), a first draft of the paper (15 marks), a paper-presentation (5 marks), and a final draft of the paper (25 marks - 5 for the abstract and 20 for the paper). More details for each of these components will be given on the course webpage, but a short explanation is given below.

> *Proposal*. (5 marks) One way to think of this assignment is as though you are submitting a 500-word abstract to a conference. In order for the organisers of the conference to accept your paper based on this abstract, you must provide sufficient detail to display the topic of your proposal, and a sense of what your main contribution will be.
>
> *First Draft*. (15 marks) The first draft of your paper should be around 2000 words and it should be on the topic that you proposed in the first part of this assignment. (Although, this is not, strictly speaking, necessary: you are allowed to change your mind about the topic of your paper between submitting the proposal and submitting the first draft of the paper.)
>
> *Paper Presentation*. (5 marks) In our final meeting, all of the graduate students will give a (very) short presentation on the topic of their research paper. They should prepare to present for no more than ten minutes, after which there will be around 5-10 minutes for Q & A. (*Note*: time allotted for presentations is tentative, and will be based on course enrolment. More details will be given in advance of the date.)
>
> *Final Draft*. (25 marks) The final draft of the long paper will be due on the last day of the exam period. The final paper should be a concise research paper of around 3000 words (no more than 3500) on a topic of your choosing; and, it should include an abstract of around 250 words (no more than 300 words). 5 marks will be allocated to the abstract, and the remaining 20 marks will be allocated to the rest of the paper. One way to think of this assignment is as though you are submitting a short research paper to the *American Philosophical Association*.

2.4 Overview (Graduate). The grading for the graduate section of this course is broken down as follows:

Weekly Reading Responses	10 %
2x 'Referee Reports'	20 %
2x 'Conference Commentaries' (10 min. Presentation)	20 %
Final Research Paper (around 3000 words)	50 %
TOTAL	**100 %**

The final paper has several components, and the grade for the final paper will be broken down as follows:

Proposal (500 words)	5 %
First Draft (2000 words)	15 %
Paper Presentation (in Class)	5 %
Abstract of the Final Paper (250 words)	5 %
Revised final Draft (3000 words)	20 %
TOTAL	**50 %**

2.5 Topic Overview

Week 1: Course Introduction

PART I: Concepts

Week 2: Deep Learning & AI Today

Week 3: The Control Problem

Week 4: The Alignment Problem

PART II: Approaches

Week 5: Technical Approaches to Alignment

Week 6: Approaches to Machine Ethics

Week 7: Utilitarian Approaches to AMA

Week 8: Deontological Approaches to AMA

Week 9: Virtue-Ethical Approaches to AMA

Week 10: *Fall Break* (No Class)

PART III: Consequences

Week 11: Social Consequences

Week 12: Explanation, Interpretation, Transparency

Week 13: Bias and Fairness

Week 14: Graduate Paper Presentations

Week 15: *Exam Period I* (No Class)

Week 16: *Exam Period II* (No Class)

2.6 Ground Rules for Discussion. These ground rules form a set of expected behaviours for conduct in discussions and lectures. They are meant to foster an intellectual atmosphere where we work together to achieve knowledge. They are also meant to ensure that discussions are spirited without devolving into argumentation and to ensure that everyone has an opportunity to be heard.

DO:

- Respect yourself and others (share your viewpoint and allow others to share theirs).
- Show respect for others by learning and using their preferred names and pronouns
- Give each other the benefit of the doubt. (Be charitable.)
- Be cautious of universal claims.
- Listen actively and attentively.
- Keep an open mind. (Expect to learn something new, or to have your views challenged by ideas, questions, and points of view different than your own.)
- Ask for clarification if you are confused.
- Challenge one another, but do so respectfully.
- Allow others (and yourself) to revise or clarify ideas and positions in light of new information.
- Critique ideas, not people.
- Take responsibility for the quality of the discussion.
- Build on one another's comments; work toward shared understanding.
- Try to always have your readings in front of you.
- If you are offended by anything said during discussion, acknowledge it immediately.

DO NOT:

- Interrupt one another—even when you are excited to respond.
- Offer opinions without supporting evidence.
- Engage in put-downs.
- Make assumptions—ask questions instead.
- Do not monopolise discussion.

If you notice patterns that are troubling or might be impeding full engagement by others, please speak to me in office or via email. Such discussions should be understood as being strictly confidential. If it is not possible to speak to me, feel free to reach out to the department chair, and academic advisor, or a trusted mentor.

2.7 Detailed Course Schedule

Note that each week contains many more readings than are typical of a philosophy course. This is because conferences venues, rather than journals, are the main form of publication in machine learning, and these papers are typically only 4, 6, or 8 pages long. On average, there are about 60 pages of reading per week.

Week 1
Sept. 8

Course Introduction

Stuart Russell. 2019. ‘If We Succeed’, Ch. 1 in *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking. 1-12.

PART I: Deep Learning, Value Alignment, and Control

Week 2
Sept. 15

Deep Learning and Artificial Intelligence Today

Cameron Buckner. 2019. ‘Deep Learning: A Philosophical Introduction’, *Philosophy Compass*. 14(10): e12625.

S. Matthew Liao. 2020. ‘A Short Introduction to the Ethics of Artificial Intelligence’ in S. Matthew Liao (ed.) *Ethics of Artificial Intelligence*. Oxford: Oxford University Press. 1-42.

Week 3
Sept. 22

The Control Problem

Nick Bostrom. 2014. ‘The Superintelligent Will’, Ch. 7 in *Superintelligence: Paths, Dangers Strategies*. Oxford: Oxford University Press. 127-139.

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

Stuart Russell. 2019. ‘Overly Intelligent AI’, Ch. 5 in *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking. 132-144.

Stuart Russell. 2019. ‘The Not-So-Great AI Debate’, Ch. 6 in *Human Compatible: Artificial Intelligence and the Problem of Control.* New York: Viking. 145-170.

Week 4
Sept. 29

The Alignment Problem

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

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.

PART II: Machine Ethics

Week 5
Oct. 6

Technical Problems for Value-Aligned AI

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané. 2016. Concrete problems in AI safety. *arXiv Pre-Print*. 1606.06565: 1-29. https://arxiv.org/abs/1606.06565.

Dylan Hadfield-Menell, and Gillian Hadfield. 2018. 'Incomplete Contracting and AI Alignment’. *arXiv Pre-Print*. 1804.04268: 1-16. https://arxiv.org/abs/1804.04268

Dylan Hadfield-Menell, Anca Dragan, Peter Abbeel, and Stuart Russell. 2016. 'Cooperative inverse reinforcement learning’. *Advances in Neural Information Processing Systems*. 1-9. https://arxiv.org/abs/1606.03137

Week 6
Oct. 13

Approaches to Machine Ethics

Colin Allen, Iva Smit, and Wendell Wallach. 2005. ‘Artificial morality: Top-down, bottom-up, and hybrid approaches’, *Ethics and Information Technology* 7(3): 149-155.

Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. ‘Implementations in Machine Ethics: A Survey’, *ACM Computing Surveys*. 53(6): 132:1-132:38.

Aimee Van Wynsberghe and Scott Robbins. 2019. ‘Critiquing the Reasons for Making Artificial Moral Agents’, *Science and Engineering Ethics* 25(3): 719-735.

Week 7
Oct. 20

Artificial Moral Agents (Utilitarian AI)

Stuart Russell. 2019. 'AI: A Different Approach', Ch. 7 in *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking. 171-183.

Stuart Russell. 2019. 'Provably Beneficial AI', Ch. 8 in *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking. 184-210.

Heather M. Roff. 2020. 'Expected Utilitarianism'. *arXiv Preprint*. 2008.07321: 1-22. https://arxiv.org/pdf/2008.07321.pdf.

Stuart Armstrong. 2015. 'Motivated Value Selection for Artificial Agents', in *Workshops at the 29th AAAI Conferences on Artificial Intelligence*. 1-9.

Week 8
Oct. 27

Artificial Moral Agents (Deontological AI)

Thomas M. Powers. 2005. 'Deontological Machine Ethics' in Michael Anderson, Susan Leigh Anderson, and Chris Armen (eds.) *AAAI Fall Symposium on Machine Ethics*. 1-6.

Thomas M. Powers. 2006. 'Prospects for a Kantian Machine', *IEEE Intelligent Systems* 21(4): 46-51

Susan Leigh Anderson and Michael Anderson (2011) 'A prima facie duty approach to machine ethics and its application to elder care', in *AAAIWS '11-12: Proceedings of the 12th AAAI Conference on Human-Robot Interaction in Elder Care*. Association for Computing Machinery. 2–7.

Derek Leben. 2017. 'A Rawlsian algorithm for autonomous vehicles', Ethics and Information Technology 19: 107–115.

Silviya Serafimova. 2020. 'Whose Morality? Which Rationality? Challenging Artificial Intelligence as a Remedy for the Lack of Moral Enhancement', *Humanities and Social Sciences Communications*. 7(119): 1-10.

Week 9
Nov. 3

Artificial Moral Agents (Virtue-Ethical AI)

Shannon Vallor. 2016. 'Virtue Ethics, Technology, and Human Flourishing', Ch. 1 in *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford: Oxford University Press. 17-34.

Shannon Vallor. 2016. 'The Case for a Global Technomoral Virtue Ethic', Ch. 2 in *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford: Oxford University Press. 35-57

Don Howard and Ioan Muntean. 2017. 'Artificial Moral Cognition: Moral Functionalism and Autonomous Moral Agency', in Thomas M. Powers (ed.) *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics*. Philosophical Studies Series, Vol. 128, Cham: Springer. 121-159.

Week 10
Nov. 10

Fall Break (No Class)

PART III: Social Value Alignment

Week 11 Nov. 17	**AI Ethics and AI Safety: Social Consequences** Miles Brundage. 2014. ‘Limitations and Risks of Machine Ethics’, *Journal of Experimental and Theoretical Artificial Intelligence*. 26(3): 355-372. Iason Gabriel and Vafa Ghazavi. 2021. ‘The Challenge of Value Alignment: from Fairer Algorithms to AI Safety’ forthcoming in *The Oxford Handbook of Digital Ethics*. 1-20. Stuart Russell. 2019. ‘Misuses of AI’, Ch. 4 in *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking. 103-131.
Week 12 Nov. 24	**Explanation, Interpretation, and Transparency** Kathleen Creel. 2020. Transparency in Complex Computational Systems. *Philosophy of Science*. 87(4): 568-589. Atoosa Kasirzadeh. 2019. Mathematical decisions and non-causal elements of explainable AI. *arXiv Preprint*, 1910.13607: 1-26. https://arxiv.org/abs/1910.13607 Adrian Erasmus, Tyler D. P. Brunet, and Eyal Fisher. 2020. ‘What Is Interpretability?’. Forthcoming in *Philosophy and Technology*. 1-30.
Week 13 Dec. 1	**Consequences of Value Misalignment: Bias and Fairness** Gabbrielle M. Johnson. 2020. ‘Algorithmic bias: on the implicit biases of social’, Forthcoming in *Synthese*. 1-21. Gabbrielle M. Johnson. Forthcoming. ‘Are Algorithms Value-Free? Feminist Theoretical Virtues in Machine Learning’. Forthcoming in *Journal of Moral Philosophy*. 1-33. Nenad Tomasev, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. 2021. ‘Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities’. *arXiv Preprint*, 2102.04257: 1-15. https://arxiv.org/abs/2102.04257.
Week 14 Dec. 8	**Graduate Paper Presentations (In Class)**
Week 15 Dec. 15	**Exam Period I**
Week 15 Dec. 22	**Exam Period II**

3.2 Disclaimer

This document is meant to serve as a record of policies and expectations for the purpose of accountability. However, in the event of circumstances beyond my control, the course content, evaluation scheme, and any other part of this syllabus are subject to change. If any changes are made, they will typically be done in consultation with the students, and you will be given advance notice of such changes.