University of Toronto
Faculty of Arts & Science
Department of Philosophy

'Artificial Intelligence and the Value-Alignment Problem'

PHL455 H1F: Seminar in Philosophy of Science
Summer 2021, May - June

1. COURSE INFORMATION

1.1 Course Times + Location

MW 12:00 – 15:00 Eastern (UTC-04:00); Online (synchronous), via Zoom

1.2 Office Hours + Location

T 12:00 – 15:00, or by appointment; Online, via Zoom

1.3 Contact Information

Dr. Travis LaCroix (He/Him), travis.lacroix@utoronto.ca

1.4 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 become 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*.

In the first part of this course, we will examine the concepts of control and value alignment to see how they are connected and what practical, scientific, ethical, and philosophical questions arise when trying to solve these problems. We will focus on both the normative and technical components of value-aligned artificial intelligence—namely, how to achieve moral agency in an artificial system. The normative component of the value alignment problem asks what values or principles (if any) we ought to encode in an artificial system; whereas, the technical component asks how we can encode these values. 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.

1.5 Course Objectives and Learning outcomes.

- Deep knowledge of contemporary 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.

1.6 Required course materials. All of the required readings for this course will be made available online through Quercus (https://q.utoronto.ca). See course schedule below for list of required readings. If you are auditing or on the waitlist, and do not have access to the course webpage, please email me.

There are three categories of readings for each meeting, which are denoted as 'Core', 'Background' ('*Back.*'), and 'Supplementary' ('*Supp.*'). *The core and background readings should be understood to be*

required, whereas the supplementary readings are optional; but they differ in terms of the time you are expected to devote to each of them, and how well you are expected to have understood them, according to the following scheme:

Core. Core readings will form the basis of our discussion in class. These should be read thoroughly and carefully, *at least* once, and they should be well understood in their entirety.

Back. Background readings will provide secondary content for our discussion of the core reading. These should be read in full once, but can be read slightly more quickly than core readings, and parts of the paper may be skimmed over. The key points of the argument, though not necessarily all of the details, should be understood.

Sup. Supplementary readings are meant to provide some additional details or perspectives for the core and background readings. *These are optional.* It will suffice to read just enough to get a sense of the main contribution of the paper—for example, by reading the abstract, introduction, and conclusion. A deep understanding of the content is not required.

Typically, there will be one core, one background, and one supplementary reading for each meeting.

I.7 Detailed Course Schedule.

Week 1		
3 May 2021 (M)		**Course Introduction**
	Core.	Syllabus
	Core.	Stuart Russell. 2019. 'If We Succeed', Ch. 1 in *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking. 1-12.
5 May 2021 (W)		**Deep Learning and Artificial Intelligence Today**
	Core.	Cameron Buckner. 2019. 'Deep Learning: A Philosophical Introduction', *Philosophy Compass*. 14(10): e12625.
	Back.	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. ***Note:*** **Skip Section 1.6**.
Reading Response(s) for Week 1 Readings Due at 23:59 EDT **Friday, May 7**		
Week 2		
10 May 2021 (M)		**The Control Problem**
	Core.	Nick Bostrom. 2014. 'The Superintelligent Will', Ch. 7 in *Superintelligence: Paths, Dangers Strategies*. Oxford: Oxford University Press. 127-139.

Cont'd…

Core.

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

Back.

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

Back.

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.

12 May 2021 (W)

The Alignment Problem

Core.

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

Back.

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.

Back.

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.

Sup.

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

Reading Response(s) for Week 2 Readings
Due at 23:59 EDT
Friday, May 14

Week 3

17 May 2021 (M)

Technical Problems for Value-Aligned AI

Core. 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.

Back. 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

Sup. 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

19 May 2021 (W)

Technical Approaches to Machine Ethics

Core. 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.

Cont'd…

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

Reading Response(s) for Week 3 Readings
Due at 23:59 EDT
Friday, May 21

Week 4

24 May 2021 (M)	**Victoria Day,** *No Class*
26 May 2021 (W)	**AI Ethics and Safety: Social Consequences**
Core.	Miles Brundage. 2014. 'Limitations and Risks of Machine Ethics', *Journal of Experimental and Theoretical Artificial Intelligence*. 26(3): 355-372.
Back.	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.
Sup.	Stuart Russell. 2019. 'Misuses of AI', Ch. 4 in *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking. 103-131.

Paper Outline
Due at 23:59 EDT
Wednesday, May 26

Reading Response(s) for Week 4 Readings
Due at 23:59 EDT
Friday, May 28

Week 5

31 May 2021 (M)	**Consequences of Misalignment I: Bias and Fairness**
Core.	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
Core.	Ben Green. 2019. '"Good" isn't Good Enough', in *Proceedings of the AI for Social Good Workshop at NeurIPS*. 1-8.
Back.	Brian Christian. 2020. 'Fairness', Ch. 2 in *The Alignment Problem*. W. W. Norton & Company. 51-81.

Cont'd…

	Sup.	Kevin W. Bowyer, Michael C. King, Walter J. Scheirer, and Kushal Vangara. 2020. 'The "Criminality from Face" Illusion', *IEEE Transactions on Technology and Society*. 1(4): 175-183
	Sup.	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.
2 June 2021 (W)		**Consequences of Misalignment I: Bias and Fairness**
	Core.	Gabbrielle M. Johnson. 2020. 'Algorithmic Bias: On the Implicit Biases of Social Technology', forthcoming in *Synthese*. 1-21.
	Back.	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.
	Back.	Gabbrielle M. Johnson. Forthcoming. 'Are Algorithms Value-Free? Feminist Theoretical Virtues in Machine Learning', forthcoming in *Journal of Moral Philosophy*. 1-33.
Reading Response(s) for Week 5 Readings Due at 23:59 EDT **Friday, June 4**		
Week 6		
7 June 2021 (M)		**Consequences of Misalignment II: Explanation, Interpretation, Transparency**
	Core.	Kathleen Creel. 2020. 'Transparency in Complex Computational Systems', *Philosophy of Science*. 87(4): 568-589.
	Back.	Adrian Erasmus, Tyler D. P. Brunet, and Eyal Fisher. 2020. 'What Is Interpretability?', forthcoming in *Philosophy and Technology*. 1-30.
9 June 2021 (W)		**Consequences of Misalignment II: Explanation, Interpretation, Transparency**
	Core.	Atoosa Kasirzadeh. 2019. Mathematical Decisions and Non-causal Elements of Explainable AI. *arXiv Preprint*, 1910.13607: 1-26. https://arxiv.org/abs/1910.13607
Reading Response(s) for Week 6 Readings Due at 23:59 EDT **Friday, June 11**		
Week 7		
14 June 2021 (M)		**In-Class Presentations** *No Additional Reading*

Final Paper
Due at 23:59 EDT
Monday, June 21

1.8 Course Delivery Information. Live, synchronous sessions will be held Mondays and Wednesdays, 12:00 – 15:00 Eastern Daylight Time (UTC-04:00). Zoom links for the Monday and Wednesday live sessions will be posted on Quercus (https://q.utoronto.ca). The official recommended technical requirements can be found here: https://www.viceprovoststudents.utoronto.ca/.

2. GRADING

2.1 Assessment Details. The breakdown of the final grade for this course is given as follows.

Weight	Description	Deadline
20%	Attendance and Participation	Ongoing
20%	Discussion Board Questions & Responses	Due: Weekly
20%	Paper Outline	Due: Wednesday, May 26
40%	Research Paper	Due: Monday, June 21
2%	Bonus Marks	

Details for each of these components are provided below.

2.1.1 Attendance and Participation. (20 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. Ten marks of the twenty will be allocated for attendance. Unless there is a documented excuse provided, timely arrival and attendance for the entire duration of the seminar will be counted as one mark per meeting. Excessive tardiness, or early departure, will count as a half-point for the day. Since we have 12 meetings in total, this means that you may have up to 2 undocumented absences without your grade being affected. The other ten marks will be similarly assigned for active participation in the seminar; active participation may involve (i) having one's camera on (for the majority of) the meeting; (ii) participating in the discussion—either verbally (unmuting your microphone) or in writing (in the chat); or, (iii) constructively engaging with the discussion questions that your peers have posted in the discussion forum. I will discuss this more on the first day of class. Your attendance and participation grade will be updated weekly, and any perceived errors should be brought to my attention as soon as possible.

2.1.2 Discussion Board Questions & Responses. (20 marks total)

> *Discussion Board Questions.* (10 marks total) Students should submit (10x) a substantive question in response to a single assigned reading. Each question submitted will be worth 1 mark and graded on a pass/fail basis. Ten passing submissions will be sufficient a full 10 marks on this portion of the assignment. A passing submission involves a thoughtful question that is raised in response to one of the readings, and which provides context for motivating why the question is being asked. A sample of what I consider to be a good submission (question plus context for asking the question) will be supplied on the course webpage. The discussion questions are due by **10:00 EDT** the morning of the meeting for which we are discussing the reading to which the question refers.

> ***For example***, A substantive question for Creel (2020), 'Transparency in Complex Computational Systems', which we are discussing in week 7, (Monday June 7), should be submitted online no later than 10:00, **Monday June 7**.

Discussion Board Responses. (10 marks total) Students should submit (5x) a post on the discussion board (online, 250-500 words) that engages with (at least) one of the readings assigned for a given seminar meeting. Each response will be worth 2 marks. It is possible to receive a pass (2 marks), a marginal pass (1 mark) or a fail (0 marks) for each of these. A rubric will be posted on the course webpage to clarify what differentiates a 'pass', a 'marginal pass' and a fail. The response should involve a substantive claim (like a thesis statement), and provide evidence for that claim. A sample of what I consider to be a good submission (full marks) will be posted on the course webpage. You may submit more than 5 substantive posts throughout the course of the semester; your best five grades will be counted toward your final mark. These responses for a given week's readings are due at the end of the week, each **Friday** by **23:59**.

> ***For example***, if you wanted to write a substantive response to Gabriel (2020), 'Artificial Intelligence, Values, and Alignment', which we are discussing in week 2, (Wednesday May 12), it should be submitted online no later than 23:59, Friday May 14.

Students may submit more than one response per week, as long as they are on different readings.

Note. As mentioned in the section on participation and attendance, constructively responding to another student's discussion board question/response will be counted toward the participation grade.

2.1.3 Paper Outline. (20 marks total) Students must submit an outline (i.e., a proposal) of around 750 words for their research paper. The paper outline must be submitted electronically no later than **Wednesday, May 26 at 23:59 EDT**. This assignment will be graded and returned to you by the end of class, on Monday, May 31. Students should plan to meet with me individually (for approximately 30 minutes) the week of May 31 to discuss their paper proposal and to go through some of the feedback they received in more detail. Two weeks before the deadline (i.e., by May 12), I will post a list of available meeting times on the course webpage. Students should sign up for a meeting slot to discuss the proposal, the feedback, and your plan for your research paper. More details will be given in class.

2.1.4 Short Research Paper. (40 marks total) A final research paper will be due on **Monday, June 21 at 23:59 EDT**. The paper should expand upon the outline that was proposed and discussed. During our final meeting (Monday, June 14), each student will give a brief (5 min.) presentation of the main argument of their paper, which will be followed by 5-10 min. Q&A. Authors may include a handout or a short slide deck for their presentation, but this is entirely optional.

> ***Note*** : Actual allotted times for presentations and Q&A will depend on course enrollment; if necessary, some presentations will be held in the second half of our second-to-last meeting, on Wednesday, June 9. A sign-up sheet will be posted on the course webpage, and more details will be given in class.

The final paper, when submitted, should include an abstract of up to 150 words and a 'synopsis' (a clean, revised version of the proposal which summarises the main components of the paper) of up to 750 words. The final paper itself should be around 3000 words, excluding footnotes and references. Further details will be given in class and on the course webpage. The final grade for the paper will be broken down as follows:

5% In-class presentation and Q&A
5% Abstract (150 words)
5% Paper synopsis (up to 750 words)
25% Body of paper (3000 words, excluding footnotes and references)

2.1.5 Bonus Marks. (2 marks total) 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 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, 'true/false', etc.) 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 day of class (23:59 EDT, **Wednesday, May 5**) 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.

2.2 Submitting Work. All your work should be submitted electronically via Quercus. The paper outline and the final paper must be in .pdf format. You are responsible for ensuring that your work is uploaded, and that the file submitted is the one that you intended to submit. (For example, it is a good practice to check the document a few minutes after you uploaded it to ensure everything is in order.)

Normally, students will be required to submit their course essays to Turnitin.com for a review of textual similarity and detection of possible plagiarism. In doing so, students will allow their essays to be included as source documents in the Turnitin.com reference database, where they will be used solely for the purpose of detecting plagiarism. The terms that apply to the University's use of the Turnitin.com service are described on the Turnitin.com web site.

3. COURSE POLICIES

3.1 Notice of video recording and sharing (download and re-use prohibited). ~~This course, including your participation, will be recorded on video and will be available to students in the course for viewing remotely and after each session. Course videos and materials belong to your instructor, the University, and/or other sources depending on the specific facts of each situation and are protected by copyright. Do not download, copy, or share any course or student materials or videos without the explicit permission of the instructor. For questions about the recording and use of videos in which you appear, please contact your instructor.~~ As per the student poll at the first meeting, the seminar discussions will *not* be recorded. [For those who were not at the first meeting, inclusion of this language on the syllabus was a legal

requirement for the possibility of recording the meetings. Whether the discussions were recorded or not is a decision which I left up to the students.]

3.2 Contact Policy. My email policy is to respond to any enquiries within two workdays 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 ('PHIL455') 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.

3.3 Late Submission Policy. Work that is submitted late, without a documented excuse, will be penalized 2% per 24-hour period after the deadline. *This policy applies only to the paper outline and the research paper. No late submissions will be accepted for the reading questions/responses.*

3.4 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.

Please note the following addenda that were agreed upon in the first class: (1) we ought to take care with the terms we use. This goes both ways: one should try to be clear, especially when using fuzzy concepts or terms with multiple meanings (especially across disciplines), but also one should take care in interpretating others' use of such terms—when in doubt, prefer to ask for clarification rather than talking past one another. (2) Acknowledge the contributions of others, for example by giving credit where credit is due. Also, it does not hurt to bolster [support] your colleagues positively: if you think someone has a really interesting idea, let them know! (3) Do be sure to respond to responses. If you are discussing a

point, and someone adds to that point, acknowledge it and respond to it (or let someone else respond to it) before going back to the original point. These rules have been collectively agreed upon by this particular group of students. While many may go without saying, it is important that they are understood to be common knowledge. This sets the precedent for how we interact with one another over the course of the next six weeks.

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.

3.5 Academic Integrity. All work submitted must be your own. Academic integrity is essential to the pursuit of learning and scholarship in a university, and to ensuring that a degree from the University of Toronto is a strong signal of each student's individual academic achievement. As a result, the University treats cases of cheating and plagiarism very seriously. The University of Toronto's Code of Behaviour on Academic Matters outlines the behaviours that constitute academic dishonesty and the processes for addressing academic offences. Potential offences include, but are not limited to:

(For papers and assignments)

1. Using someone else's ideas or words without appropriate acknowledgement;
2. Submitting your own work in more than one course without the permission of the instructor;
3. Making up sources or facts;
4. Obtaining or providing unauthorized assistance on any assignment

With regard to remote learning and online courses, students are expected to adhere to the Code of Behaviour on Academic Matters regardless of the course delivery method. By offering students the opportunity to learn remotely, it is expected that students will maintain the same academic honesty and integrity that they would in a classroom setting. Potential academic offences in a digital context include, but are not limited to:

1. Accessing unauthorized resources (search engines, chat rooms, Reddit, etc.) for assessments.
2. Using technological aids (e.g. software) beyond what is listed as permitted in an assessment.
3. Posting test, essay, or exam questions to message boards or social media.
4. Creating, accessing, and sharing assessment questions and answers in virtual "course groups."
5. Working collaboratively, in-person or online, with others on assessments that are expected to be completed individually.

All suspected cases of academic dishonesty will be investigated following procedures outlined in the Code of Behaviour on Academic Matters. If you have questions or concerns about what constitutes appropriate academic behaviour or appropriate research and citation methods, you are expected to seek out additional information on academic integrity from your instructor or from other institutional resources. More information on how **not** to plagiarise can be found here: https://advice.writing.utoronto.ca/using-sources/how-not-to-plagiarize/

5. UNIVERSITY RESOURCES

Me. Part of my job is to help you succeed, so please don't hesitate to ask me for help. I'm always happy to talk with you about anything related to the course.

Additional resources can be found at the following location:
https://www.utoronto.ca/utogether/resources-for-students