Ethics in AI Syllabus
Liam Kofi Bright

Course intent: there is no denying that the automation of much of our working lives and governance structure is a matter of great public concern. As such it would be to a citizen's advantage that they should be able to assess and discuss these issues in an informed and intelligent manner. This course is designed to facilitate that.

The first half focuses on what I think of as the foundational points of the discussion. First, what is morally and politically at stake in the wave of automation we are now undergoing. Second, given that so much turns on new machines' superior ability to reason about various issues, just what sort of epistemic capacities can we reasonably expect from the automata. Once students are grounded in these issues (including in how the two sets of concerns interrelate) we will examine a series of issues that have aroused public concern.

There are of course many more issues than I could possibly cover in one course. I have chosen to give a week each to stakeholder-transparency, medical uses, labour rights, privacy, and aligning AI values with designer values. No suggestion is intended that these are all that matters. Rather the hope is that by discussing a series of cases in some depth it will give students the practice at applying the foundational knowledge from the first half to a variety of contexts and thus help them extend their reasoning to other issues we did not have time to discuss.

Teaching Structure: the course is designed around my typical mode of teaching a 10 week course. I have one "primary reading" which I lecture on. This I typically use to introduce a broad topic and any arguments or concepts I think are of especial interest. Then I have a "secondary reading" which I lead a discussion based seminar on. These are supplemented by a set of optional readings which I will not insist students read but which greatly enhance the experience if they do.

Re assessment my desiderata for this course is largely just that students have a sophisticated understanding of the issues the course deals with. As such I would assess with a long form essay on a topic of their choice based on any of the 10 weeks' material. Students would write a draft essay or essay plan and submit that to me for initial feedback. They would then be graded on the rewrite of that essay at the end of the course. Part of the desiderata would be that the student adequately responds to feedback, since I think understanding these issues requires being able to intelligently partake in back and forth. So included within the final essay should be a short appendix (say, no more than 800 words) regarding how they responded to feedback and why they made the choices they did.

Prerequisites: the combination of skills required for this course means that it is probably appropriate as an upper level undergraduate or masters level course, and while I would not insist on it some prior experience of moral/political philosophy and logic/statistics would be useful.

Week 1: Ethical Foundations I - Bias:

Primary reading: *Algorithmic bias* - Sina Fazelpour and David Danks https://compass.onlinelibrary.wiley.com/doi/10.1111/phc3.12760?af=R

Secondary reading: *Are Algorithms Value-Free?* - Gabrielle M. Johnson https://www.gmjohnson.com/uploads/5/2/5/1/52514005/are_algorithms_value_free_.pdf

Optional readings: *Conscientious Classification.* Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta https://www.liebertpub.com/doi/abs/10.1089/big.2016.0048

What is Bias? - Gabrielle M. Johnson https://academic.oup.com/mind/article-abstract/129/516/1193/5841111?redirectedFrom=fulltext&login=false

Learning Optimal Fair Policies. - Razieh Nabi, Daniel Malinsky, and Ilya Shpitser http://proceedings.mlr.press/v97/nabi19a/nabi19a.pdf

Science and Human Values - Carl Hempel https://philpapers.org/rec/HEMSAH

Inductive risk and Values in Science - Heather Douglas https://www.jstor.org/stable/188707#metadata_info_tab_contents

Du Bois Democratic Defence of the Value Free Ideal - Liam Kofi Bright https://link.springer.com/article/10.1007/s11229-017-1333-z

Course Narrative Notes: starting the course with a week on bias is an attempt to meet students where they're at. Talk about bias or prejudice in algorithmic decision making is a huge part of public conversation. As such an early demonstration of the class' ability to clarify will hopefully inspire confidence in students that there's a literature which sheds light on and elevates such discussion.

The particular readings chosen are supposed to guide students in the following way. The introductory reading is from a survey article explicitly designed to help students gain precision and clarity on different senses of "bias". This will then help discussion on the Johnson article wherein it is discussed just what value-freedom, one interpretation of freedom from bias, might mean in this context. The optional readings provide greater philosophical depth of analysis on the notion of bias and the debate around value freedom, as well as provide a couple of examples of computer scientists and others attempting to actually implement anti-bias fairness concerns in their work.

Students should come away with a sense that through analysis they can clarify the terms of debate and refine the arguments they wish to make. What is more, they should have greater first order knowledge of where the morally significant choice points in algorithm design are re bias and fairness.

Week 2: Ethical Foundations II - Justice

Primary reading: *Algorithmic injustice* - Abeba Birhane https://www.sciencedirect.com/science/article/pii/S2666389921000155

Secondary reading: *Data Owning Democracy or Digital Socialism?* - James Muldoon https://www.tandfonline.com/doi/full/10.1080/13698230.2022.2120737

Optional readings: *Justice Beyond Utility* - Lily Hu https://dl.acm.org/doi/abs/10.1145/3278721.3278798

Algorithmic Fairness and the Situated Dynamics of Justice - Sina Fazelpour, Zachary C. Lipton and David Danks https://philpapers.org/rec/FAZAFA

Artificial Intelligence in a Structurally Unjust Society - Ting-An Lin and Po-Hsuan Cameron Chen https://ojs.lib.uwo.ca/index.php/fpq/article/view/14191

The Steep Costs of Capture - Meredith Whittaker https://dl.acm.org/doi/fullHtml/10.1145/3488666

Algorithmic Racial Discrimination - Alysha Kassam and Patricia Marino https://ojs.lib.uwo.ca/index.php/fpq/article/view/14275

Data Science as Political Action - Ben Green https://ieeexplore.ieee.org/abstract/document/9684742

The Society of Algorithms - Jenna Burrell and Marion Fourcade https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-090820-020800

Course Narrative Notes: moving from bias to justice should help ensure students do not fall prey to one of the defects of public discussion — a tendency to discuss algorithmic decisions on the basis of particular decisions and their consequences for particular individuals. Of course this is important, but it can neglect or obscure the total social effects of deciding to make decisions one way rather than another.

The particular readings chosen should guide the students as followers. The first reading by Birhane introduces the general issue of algorithmic justice when considered at the societal perspective while arguing for an interesting perspective thereon. The second reading is then deliberately chosen to be on an issue that one could not even see arising if one maintained the focus on individual level decisions and their consequences for immediate participants. The optional readings are all straightforward elaborations of points made in the main readings, or introducing and defending new and alternative perspectives on the topic.

Students should now be able to apply their more sophisticated understandings of fairness in individual decisions to arguments about broader social impact. Make sure to include some time in seminar discussion dedicated to this.

Week 3: Explanatory Desiderata I - Accuracy

Primary reading: *Inherent Trade-Offs in the Fair Determination of Risk Scores* - Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan https://arxiv.org/abs/1609.05807

Secondary reading: *Robustness in machine learning explanations: does it matter?* - Leif Hanxox-Li https://dl.acm.org/doi/abs/10.1145/3351095.3372836

Optional readings: *The games we play* - Abeba Birhane and David J. T. Sumpter https://arxiv.org/abs/2205.08922

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification - Joy Buolamwini and Timnit Gebru http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Dissecting racial bias in an algorithm used to manage the health of populations - Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan https://www.science.org/doi/10.1126/science.aax2342

The Independence Thesis - Conor Mayo-Wilson, Kevin J. S. Zollman and David Danks https://www.jstor.org/stable/10.1086/661777#metadata_info_tab_contents

Escaping the Impossibility of Fairness - Ben Green https://link.springer.com/article/10.1007/s13347-022-00584-6

Course Narrative Notes: by now students should be fired up. But they must eat their vegetables! So before we get too ahead of ourselves we take some time to ensure we are grounded in just what automation is actually doing. Almost all of the uses of algorithmic automation we are concerned with in this course involve, at some point or another, the use of some sort of machine learning process to make a prediction about a person or process. We must therefore ask ourselves — what sorts of things are being predicted and what would it mean to make a good prediction in these consequences?

The opening paper is a famous illustration that there are apparently deep conflicts involved in getting all the things right that we should like to get right. Then the second paper should complicate students' understanding even further by making it clear that the notion of "getting it right" in these circumstances is far from philosophically (or just practically, in application) simple. Two of the optional readings provide students with context re how concerns about inaccuracy arise in application. *Games* supplements the Hancox-Li, *Independence* shows the individual vs social trade off, and *Escaping* foreshadows future discussion.

This week should curb students' enthusiasms a bit. If they had been getting the impression that armed with conceptual sophistication they hence just need a righteous crusade to ensure in future we get things right, they should now be a bit more diffident and introspective.

Week 4: Explanatory Desiderata II - Causal Inference

Primary reading: *Variation Semantics* - Laurenz Hudetz and Neil Crawford http://philsci-archive.pitt.edu/20626/

Secondary reading: *What is "Race" in Algorithmic Discrimination on the Basis of Race?* - Lily Hu https://scholar.harvard.edu/files/lilyhu/files/what_is_race.pdf

Optional readings: *Causal Discovery Algorithms.* - Daniel Malinsky and David Danks http://www.dmalinsky.com/uploads/4/9/9/3/49933045/philcomp18_published.pdf

On the Explanatory Depth and Pragmatic Value of Coarse-Grained, Probabilistic, Causal Explanations. - David Kinney https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/on-the-explanatory-depth-and-pragmatic-value-of-coarsegrained-probabilistic-causal-explanations/C286C1926340B569DB85DB565445DFE0

Evaluations of Causal Claims Reflect a Trade-Off Between Informativeness and Compression - David Kinney and Tania Lombrozio http://davidbkinney.com/Causal_Evaluation_Camera_Ready.pdf

Causal and Evidential Conditionals - Mario Günther https://link.springer.com/article/10.1007/s11023-022-09606-w

"But What Are You Really?" On the Metaphysics of Race - Charles Mills https://philpapers.org/archive/MILBWA

Course Narrative Notes: continuing on the vegetable eating theme this is probably the most technically challenging of the weeks. But, as the first reading will make clear, mere accuracy is often not enough. What we would like for both our explanatory and policy making purposes is very often causal information. So how can we get that and what special issues arise when we face such question in the context of algorithmic decision making?

The first reading introduces the concept by nicely talking through a particular proposal about what sort of judgements we want to be able to make accurately through the analysis of automated decisions. And the second discussion piece then further complicates the (standard) interventionist model, which was introduced in the Hudetz & Crawford, by showing that when combined with the sort of reasoning about demographic groups we are often concerned with it has rather counter-intuitive consequences. The further readings give an introduction to some technical considerations regarding causal inference, as well as provide some background to the social constructivist ideas of race that Hu was presupposing knowledge of.

By now students should be thoroughly perplexed.

Week 5: the Good vs the True?

Primary reading: *The Bias Dilemma* - Oisín Deery and Katherine Bailey https://ojs.lib.uwo.ca/index.php/fpq/article/view/14292

Secondary reading: *On the Advantages of Distinguishing Between Predictive and Allocative Fairness in Algorithmic Decision-Making* - Fabian Beigang - https://link.springer.com/article/10.1007/s11023-022-09615-9

Optional readings: *On the implicit biases of social technology.* Gabbrielle M. Johnson https://link.springer.com/article/10.1007/s11229-020-02696-y

Connecting ethics and epistemology of AI - Federica Russo, Eric Schliesser, and Jean H.M. Wagemans http://philsci-archive.pitt.edu/21528/

On the Epistemic Costs of Implicit Bias - Tamar Szabó Gendler https://www.jstor.org/stable/41487720#metadata_info_tab_contents

Moral Obligation and Epistemic Risk - Boris Babic and Zoë A. Johnson King https://static1.squarespace.com/static/5930a7361e5b6ce07837229b/t/5d7690799a6c184e2ea07cbb/1568051322825/Moral+Obligation+and+Epistemic+Risk+-+December+2018.pdf

Normativity, Epistemic Rationality, and Noisy Statistical Evidence - Boris Babic, Anil Gaba, Ilia Tsetlin, and Robert L. Winkle https://borisbabic.com/research/noisy_stereotypes_march2021.pdf

The Myth in the Methodology - Ben Green and Lily Hu *https://www.semanticscholar.org/paper/The-Myth-in-the-Methodology%3A-Towards-a-of-Fairness-Green-Hu/2a00d745958b8916e8044df1b68d11cdf6fcc000*

Course Narrative Notes: this week is triumphant summation. Students have seen that significant and contentious evaluations can be buried in apparently neutral algorithms or implementation. They also have seen that specifying just what it is one wants from automated decision making must be done carefully if there is to be any hope of anything more than a morally compromised muddle.

Here students encounter the concern that doing well re all previous concerns may be impossible! It's a somewhat overblown concern. But it is a common worry so worth facing head on. The opening essay plainly states the case for a clash between epistemic and ethical desiderata. The second essay diffuses the concern. Secondary readings provide background on pertinent philosophical debates.

Psychologically this week is important. The hope is to imbue students with a sense of optimism. We wracked up difficulty and paradox but found them resolvable, or at least capable of significant progress! We are not just doomed to bad outcomes or powerless before politically pernicious machines and their owners. We can do better through careful analysis and the willingness to apply our knowledge.

Week 6: Transparency

Primary reading: *Transparency in Complex Computational Systems* - Kathleen Creel https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/transparency-in-complex-computational-systems/4DB040EB28172CADF5F2858B62D0952C

Secondary reading: *Algorithmic and Human Decision Making* - Mario Günther and Atoosa Kasirzadeh https://www.mario-guenther.com/_files/ugd/70b9dd_ff087ae509034fb9b126dcf783182457.pdf

Optional readings: *The Right to an Explanation* - Kate Vredenburgh https://static1.squarespace.com/static/5baaf2ad01232c0e635e3a93/t/614f30f462ced34a714c4db2/1638790800365/+The+Right+to+Explanation.pdf

Informatics of the Oppressed - Rodrigo Ochigame https://logicmag.io/care/informatics-of-the-oppressed/

Epistemic values in feature importance methods - Leif Hancox-Li and I. Elizabeth Kumar https://dl.acm.org/doi/10.1145/3442188.3445943

Fake News and Partisan Epistemology - Regina Rini https://kiej.georgetown.edu/fake-news-partisan-epistemology/

Stop Talking About Fake News! - Joshua Habgood-Coote https://philpapers.org/rec/HABSTA

Automated Trouble - Florian Saurwein and Charlotte Spencer-Smith https://www.cogitatiopress.com/mediaandcommunication/article/view/4062

Modelling How False Beliefs Spread - Cailin O'Connor and James Owen Weatherall http://cailinoconnor.com/wp-content/uploads/2020/09/A30-Modeling-How-False-Beliefs-Spread.pdf

Course Narrative Notes: now we move from the foundational section of the course to the varied applications. The first such concerns transparency, the degree to which users or affected stakeholders, can examine and comprehend the basis on which decisions are being made.

I am not fully confident in the presentation order I have decided on here and in revising the course I would expect to return to this. Presently it is structured such that the introductory readings are high level philosophical discussion of what transparency means and why we should care about it. Two of the secondary readings provide further background, then the rest are an introduction to a topic (social media algorithms and the spread of misinformation) where people find lack of transparency most concerning. It is possible it would be better to reverse this, and have the week focussed on misinformation but include readings on transparency as background.

Week 7: Labour Rights

Primary reading: *Freedom at Work* - Kate Vredenburgh https://www.cambridge.org/core/journals/canadian-journal-of-philosophy/article/freedom-at-work-understanding-alienation-and-the-aidriven-workplace/7C2D79FB942961B33C2C9D01203D2D3F

Secondary reading: *Algorithmic Domination in the Gig Economy*. - James Muldoon and Paul Raekstad https://journals.sagepub.com/doi/full/10.1177/14748851221082078

Optional readings: *A Short-term Intervention for Long-term Fairness in the Labor Market* - Lily Hu and Yiling Chen. https://dl.acm.org/doi/abs/10.1145/3178876.3186044

Structural domination in the labor market - Lillian Cicerchia https://philpapers.org/rec/CICSDI

Digital innovation and the fourth industrial revolution - Loris Caruso https://link.springer.com/article/10.1007/s00146-017-0736-1

The Algorithmic Leviathan - Kathleen Creel and Deborah Hellman https://www.cambridge.org/core/journals/canadian-journal-of-philosophy/article/algorithmic-leviathan-arbitrariness-fairness-and-opportunity-in-algorithmic-decisionmaking-systems/3AA0ECA77F8622488E9DB0834287215B

Picking on the Same Person - Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang https://arxiv.org/abs/2211.13972

Accessible Crowdwork? - Kathryn Zyskowski, Meredith Ringel Morris, Jeffrey P. Bigham, Mary L. Gray, Shaun K. Kane https://dl.acm.org/doi/abs/10.1145/2675133.2675158

Course Narrative Notes: the next week's application concerns labour rights, with a focus on workplace democracy. The previous issue (transparency) is a nice example of the relationship between epistemic concerns and considerations of individual fairness arising in practice. This issue on the other hand is better suited to highlighting how it is that society wide considerations interact with the application of automated decision. Who is empowered, how, and what potentialities may exist for doing better can naturally be raised here.

The main readings focus on workplace democracy and exploitation. But it is worth drawing attention to the two essays on monoculture, which considers the experience and perspective of job seekers. What is relevant here is that a hitherto unknown form of discrimination is made possible by automation. The Zyskowski *et al* also provides a useful alternate perspective by highlighting opportunities for greater inclusion presented by automating new features of workplace automation.

Week 8: Privacy

Primary reading: *A modern Pascal's wager for mass electronic surveillance.* - David Danks https://static1.squarespace.com/static/5f6d0320212a261d8716949f/t/6213191469077794d4dba3724/1645418773886/Telos-PascalsWager-Pub.pdf

Secondary reading: *The Surveillance Society* - Oscar H. Handy Jnr. https://academic.oup.com/joc/article-abstract/39/3/61/4210548

Optional readings: *Philosophy of Privacy and Digital Life* - Anita L. Allen https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4022657

Privacy and Paternalism: The Ethics of Student Data Collection - Kathleen Creel and Tara Dixit - https://mit-serc.pubpub.org/pub/privacy-and-paternalism/release/2

Recommender systems and their ethical challenges - Silvia Milano, Mariarosaria Taddeo and Luciano Floridi https://link.springer.com/article/10.1007/s00146-020-00950-y

What's Wrong with Automated Influence? - Claire Benn and Seth Lazar https://www.cambridge.org/core/journals/canadian-journal-of-philosophy/article/whats-wrong-with-automated-influence/98F5E24BEADE585050B773D2CBEB1F39

Protecting data privacy is key to a smart energy future - Carissa Véliz and Philipp Grunewald https://www.nature.com/articles/s41560-018-0203-3.epdf

Course Narrative Notes: another very widely discussed issue with the role of ever more automation, the increased capacity for spying upon, and manipulation of, individuals based on extensive data concerning their activities. The readings are chosen to try and draw attention to both the social consequences of this and also the way it can feel for individuals enmeshed in this system.

The placement of this topic after both Transparency and Labour Rights is quite deliberate. Being under constant surveillance in our capacity as citizens, consumers, and workers, may affect any of these spheres differently, as well as having a compounding joint effect. The Creel and Trixit paper is included to facilitate one such discussion on a matter that should be intimately familiar to all students, though one may substitute in one's preferred alternative. The Véliz and Grunewald is another way in to giving people a sense of what is at stake here.

Starting with the Danks is calculated to get people thinking about why anyone might think any of this desirable at all. This is a useful prompt for getting students to think about how decisions are made regarding the deployment of new technology, and so can naturally be used to link the discussion back to previous weeks regarding issues of social justice, and in particular Muldoon's reading on the distribution of control of automated assets.

Week 9: Medical Decisions

Primary Reading: *Impacts on Trust of Healthcare AI* - Emily LaRosa and David Danks https://static1.squarespace.com/static/5f6d0320212a261d8716949f/t/62131a243cebf6617301c7e7/1645419044243/aies132fp-larosa-pub.pdf

Secondary Reading: *Beware explanations from AI in health care* - Boris Babic, Sara Gerke, Theodorus Evgeniou, and I. Glenn Cohen https://www.science.org/doi/10.1126/science.abg1834

Optional Readings: *Difficult trade-offs in response to COVID-19* - Norheim, Ole F., Joelle M. Abi-Rached, Liam Kofi Bright, Kristine Bærøe, Octávio LM Ferraz, Siri Gloppen, and Alex Voorhoeve. https://www.nature.com/articles/s41591-020-01204-6

Reading Race - Banerjee et al https://arxiv.org/abs/2107.10356

Value judgments in a COVID-19 vaccination model. - Stephanie Harvard, Eric Winsberg, John Symons, Amin Adibia https://www.sciencedirect.com/science/article/pii/S0277953621006559

Patient Autonomy and Withholding Information. - Mellisa Rees. https://drive.google.com/file/d/14dXAVM4EwPR9sOnpcYDl-XMnoXhpfnns/view

Clinical decisions using AI must consider patient values. - Jonathan Birch, Kathleen A. Creel, Abhinav K. Jha & Anya Plutynski https://www.nature.com/articles/s41591-021-01624-y

Public deliberation and the fact of expertise - Cathrine Holst and Anders Molander https://www.tandfonline.com/doi/full/10.1080/02691728.2017.1317865

Course Narrative Notes: I was almost tempted to make this the last issue discussed, Here we want both accurate and causally explanatory information, issues of privacy and transparency are paramount, individual vs social level optimisation can easily come apart, and issues of expertise and democracy are straight to the fore. It's all here, there is a rich prior philosophical literature that one can draw from to illustrate any point one considers noteworthy, and it is on a matter of such obvious urgent importance that no one can fail to be at least somewhat interested.

The readings are organised around the theme of expertise vs democratic oversight. This is certainly not the only angle by which to introduce this and teachers should focus on whatever works best for them. But in my case one can roughly think of the Babic *et al* and the Rees as making the case for expert judgement, while the LaRosa & Danks, COVID papers, and the Birch *et al* make the case for more democratic input. Holst and Molander chart a middle course, while the Banerjee *et al* illustrates the sort of worries that motivate LaRosa and Danks while also connecting this to issues of fairness, accuracy, and transparency.

Week 10: AI Governance

Primary Reading: *Open Democracy and Digital Technologies* - Hélène Landemore https://pacscenter.stanford.edu/wp-content/uploads/2020/12/Chapte-2_9780226748436_1stPages_Mktg-2-1.pdf

Secondary Reading: *Building Governance in Online Communities* - Amy X. Zhang, Grant Hugh, Michael S. Bernstein https://dl.acm.org/doi/10.1145/3379337.3415858

Optional Reading: *Ethics as an Escape from Regulation* - Ben Wagner https://www.jstor.org/stable/j.ctvhrd092.18#metadata_info_tab_contents

Risk Imposition by Artificial Agents - Johanna Thoma https://johannathoma.files.wordpress.com/2021/02/moral-proxy-problem-feb-2021.pdf

Beyond Ethics Councils - Jana Mišić https://dig.watch/event/14th-internet-governance-forum/beyond-ethics-councils-how-really-do-ai-governance

The Epistemology of Democracy - Elizabeth Anderson https://muse.jhu.edu/article/209431

Of The Ruling of Men - WEB Du Bois http://www.webdubois.org/lectures/DuBois;OfTheRulingOfMen.html

Who Owns Your Data? - Stephen Black https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3987369

Artificial Intelligence, Values, and Alignment - Iason Gabriel https://link.springer.com/article/10.1007/s11023-020-09539-2

Council Democracy - James Muldoon https://ore.exeter.ac.uk/repository/bitstream/handle/10871/31860/Council%20Democracy%20Introduction.pdf?sequence=3

Course Narrative Notes: in the last week we arrive at the fundamental question which shall decide how all the other issues are addressed — who is in charge?

The course readings here are designed to achieve two things. First, they introduce students to democratic digital governance. Since I assume this is an unfamiliar (if only because as yet largely unexperienced) the readings focus here. The optional readings give background on democratic theory so students get more context.

Second, they provide a critique of the present governance model. Private companies control all, so they decide what the automata shall do. In this regime ethical reasoning is actually used to avoid accountability. Students should see this was not an inevitability but has come about due to some canny political operations from those who presently own the means of digital production.

Unused Week: Alignment

Primary reading: *Ethical Issues in Advanced Artificial Intelligence* - Nick Bostrom https://nickbostrom.com/ethics/ai

Secondary reading: *Risk Imposition by Artificial Agents* - Johanna Thoma https://johannathoma.files.wordpress.com/2021/02/moral-proxy-problem-feb-2021.pdf

Optional readings: *X-Risk Analysis for AI Research* - Dan Hendrycks, Mantas Mazeika https://arxiv.org/abs/2206.05862

Algorithmic Decision-Making and the Control Problem - John Zerilli, Alistair Knott, James Maclaurin and Colin Gavaghan https://link.springer.com/article/10.1007/s11023-019-09513-7

Impossibility and Uncertainty Theorems in AI Value Alignment - Peter Eckersley https://arxiv.org/pdf/1901.00064.pdf

The ethics of artificial intelligence - Nick Bostrom, Eliezer Yudkowsky http://faculty.smcm.edu/acjamieson/s13/artificialintelligence.pdf

The case for strong longtermism - Hilary Greaves and William MacAskill https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf

Stop the Robot Apocalypse - Amia Srinivasan https://discovery.ucl.ac.uk/id/eprint/1471888/1/Srinivasan%20MacAskill%20review%20final.pdf

Artificial Intelligence, Values, and Alignment - Iason Gabriel https://link.springer.com/article/10.1007/s11023-020-09539-2

Course Narrative Notes: students can have little a Skynet, as a treat.