COMPSCI294, Section 166: Foundations for Beneficial AI
Spring 2020

Stuart Russell (russell@berkeley.edu) **, Lara Buchak** (buchak@berkeley.edu) , **Wesley Holliday** (weshol liday@berkeley.edu), **Shachar Kariv** (kariv@berkeley.edu)
Course GSI: Tom Gilbert (tg340@berkeley.edu)

Class Time and Location: Mondays 2-4pm, 103 Moffitt

Dates: January 27, 2020 – April 27, 2020

Note: Attendance at first session is mandatory for all waitlisted students.

This interdisciplinary course examines the application of ideas from philosophy and economics to the problem of ensuring that increasingly intelligent AI systems remain beneficial to humans. Solving this problem requires designing AI systems whose objective is to satisfy human preferences while remaining necessarily uncertain as to what those preferences are. The course will study issues arising when applying these principles to make decisions on behalf of multiple humans and real (rather than idealized) humans. Topics include utility theory, bounded rationality, utilitarianism, altruism, interpersonal comparisons of utility, preference learning, plasticity, risk, social choice, and inequality. Permission of the graduate student instructor required. 3 units.

Prerequisites: There are no formal prerequisites for the course, but it is appropriate only for dedicated students with a background in computer science, economic thought, and/or philosophical reasoning.

Course Requirements: Class participation is required. In addition to undertaking weekly reading responses, students will be expected to produce short memos at the conclusion of the second and third course modules, as well as a final collaborative paper at the course's conclusion.

Office Location and Hours:
Stuart Russell: 4-6pm Mondays, 8040 Berkeley Way West
Lara Buchak: 2-4pm Tuesdays, via Skype + by appointment
Wes Holliday: 2-4pm Tuesdays, 246 Moses
Shachar Kariv: By appointment
Tom Gilbert: Fridays 1-2 PM in Berkeley Way West (ground floor)

SYLLABUS DATE: February 7, 2020

Learning goals: Students will read papers from the literature in AI, philosophy, and economics and will work in interdisciplinary teams to develop substantial analyses in one or more topic areas. High-level goals of the course include creating a broad intellectual community around the question of safe and beneficial AI; further refining the idea that an AI system might be beneficial for humanity (and potentially applying it to other systems such as corporations and

governments); and contributing to national and international guidelines for AI. More specific learning outcomes will be outlined in the first (mandatory) class session.

Expectations and Evaluation: Students should welcome the opportunity to engage critical and controversial perspectives that may challenge their previously held views. Respectful, prepared, and open-minded class participation is a must for making sense of these materials.

Course Format: Professors will typically lecture for 80 minutes total (or 40 minutes for a dual lecture). This will be split in two so that students can form discussion groups in response to the first lecture, evaluating its themes and articulating questions to be posed before the second lecture. Total minute breakdown is 40 (lecture 1) + 30 (discussion) + 40 (lecture 2).

Professor Responsibilities: Professors will provide all lecture content and the main required readings for the course. They will also lead the evaluation of written work and are best prepared to answer substantive questions about the readings, as well as summarize the present state of specific domain literatures (AI, economics, philosophy).

GSI Responsibilities: This position fills two roles: 1) handling students' administrative problems, including the communication of course expectations, finalizing grading rubrics, posting lecture notes, securing audio/visual equipment as needed, and collecting student feedback on the course design; 2) nurturing an interdisciplinary community among course participants, including the connection of students with particular professors and other students to advance their collaborative work. GSI office hours are a good opportunity to discuss project ideas, general interests, and learn about complementary readings beyond required course materials.

Attendance Policy: Attendance at the first class is mandatory for enrollment or for coming off the waitlist. There is a reading assignment for the first session, available on the Berkeley bCourses website. Students will be expected to come to class every session, except for illness or other personal exigencies, and should be ready to participate, having done the reading. The point of this seminar is to build community; class readings and regular attendance are ultimately different means of working to understand and engage with cross-disciplinary perspectives.

Evaluation:

1. **Weekly assignments** should be one double-spaced page posted to BCourses, summarizing the readings' key ideas and a personal question or two the student has in response. Alternatively, professors will assign specific reading questions or problem sets for students to answer, usually one week in advance. Students may also respond to each others' posts. Responses will be graded by the GSI, typically either as full-credit or a zero (i.e. it was/wasn't done). These are due by 10:30am on Mondays for the readings assigned later that day. (25% of the grade)
2. **Short memos** should be 3-4 pages long, including a minor literature review and a position review on some selected readings. Professors will provide comments on these and students should discuss with them in office hours for additional feedback. The memos will conclude the second and third themes. (25% of the grade)
3. **Final paper.** The final papers will cover one key question of the course. Students will be divided in groups of three or four, with at least two different disciplines represented. The paper should aim to be 'publishable' or submitted to an academic conference. Students will have the

opportunity to reflect on the outlines and the themes covered during office hours and are encouraged to discuss their ideas with professors and the GSI throughout the course. (More detailed rubric to be provided later)

Late Policy: Smaller assignments (i.e. weekly writing responses) will lose a half point (out of 2 total possible) if turned in after lecture on the day they are due; if turned in one day late or more they will lose a full point. Larger assignments (i.e. memos) will lose one full point (approximately one letter grade) for each day they are late. Significantly late final papers will not be graded and may result in a failing or incomplete grade for the course. All coursework must be completed by the final due date in order to receive a passing grade.

Readings: Unless otherwise specified below, all readings will be made available through the bCourses website or are readily available via arXiv or Berkeley authentication. Readings listed below are subject to change but ample notice will be given of any change.

Laptop policy: You may use a laptop during class for access to readings but you are strongly encouraged to take handwritten notes so that your screen time does not detract from the class discussion. Please keep laptops closed when not in immediate use. No cell phones or other electronic devices are allowed during class time without prior permission.

Academic Integrity: While this syllabus strives to be composed in the spirit of the UC Berkeley Honor Code and carried out in deed in our meetings, please take some time to re-acquaint yourself with it by visiting http://asuc.org/honorcode/index.php. **Plagiarism and cheating are absolutely unacceptable** in this and any other course at this University. While this seminar will encourage and emphasize group work and collaboration, the materials you turn in should be the result of your own collaborative and properly attributed work.

Students with Disabilities: If you require disability-related accommodations in this class, please meet with us and furnish a hard copy of your DSP letter by the second week of class (February 3). For more information on services available to students with disabilities please visit the DSP website (http://dsp.berkeley.edu) or the Disabled Students' Program Office located at 260 César Chávez Student Center No. 4250.

1. How do computer scientists, economists and philosophers think about Humans and AI?

<u>1/27 Session 1:</u> How do computer scientists think about AI? Presentation of AI by Prof. Russell

Stuart Russell, *Human Compatible* (Viking/Penguin, 2019), particularly Chapters 1 and 2 for today; 5, 7, 8, 9 to prepare for future lectures; and (particularly for those who have not taken an AI course), appendices A, B, C, D for background and perspectives on AI.

Russell and Norvig. *Artificial Intelligence: A Modern Approach*. Chapters 1 and 2

Suggested Reading:
Wiener, Norbert. "Some moral and technical consequences of automation." *Science* 131.3410 (1960): 1355-1358.

2/3 Session 2: How do economists think about AI? Presentation by Prof. Kariv

Ariel Rubinstein. *Lecture Notes in Microeconomic Theory*. Chapters 1, 2, 3, 7.

Suggested Reading:
Choi, Syngjoo, et al. "Who is (more) rational?." *American Economic Review* 104.6 (2014): 1518-50.

Assignment #1 (from Rubinstein):

- Problem Set 1 Question 6 (p.25)

Choose between:

- Problem Set 2 Question 1 (p. 35)
- Problem Set 3 Question 1 (p.58)

2/10 Session 3: How do philosophers think about AI? Presentation by Prof. Buchak and Prof. Holliday

- the philosophy of preferences + preference aggregation

Dreier, “Rational Preference: Decision Theory as a Theory of Practical Rationality"

Gibbard, “Interpersonal comparisons: preference, good, and the intrinsic reward of a life"

Suggested Reading:
Chang, Ruth. "The possibility of parity." *Ethics* 112.4 (2002): 659-688.

Hildebrandt, Mireille. "Privacy as protection of the incomputable self: From agnostic to agonistic machine learning." *Theoretical Inquiries in Law* 20.1 (2019): 83-121.

Kuipers, Benjamin. "Perspectives on Ethics of AI: Computer Science." (2019).

Schiffer, Stephen. "The epistemic theory of vagueness." *Philosophical Perspectives* 13 (1999): 481-503.

Homework questions:

1. According to Dreier, what problem does fine individuation pose to decision theory as a theory of rational preferences? How does he resolve this problem? How much of a problem does fine individuation pose for using decision theory as a foundation for preferences in AI?

2. Gibbard argues that “The switch to preference satisfaction as a standard of personal welfare

has been a mistake" (p. 168). Succinctly summarize his main arguments (bullet points are fine). Do you agree with these arguments? If not, briefly explain why.

2. How do we model what single humans want?

2/24 Session 4: The nature and rationality of preferences - Prof. Buchak and Prof. Kariv

Einav & Levin, Economics in the age of big data, Science 2014: Vol. 346, Issue 6210.

Gelman, Kariv, Shapiro, Silverman & Tadelis, Harnessing naturally occurring data to measure the response of spending to income, Science 2014: Vol. 345, Issue 6193.

McClennen, Edward F. "Sure-thing doubts." *Foundations of utility and risk theory with applications*. Springer, Dordrecht, 1983. 117-136.

Weekly Assignment:
- write a short 'referee' report on the Einav and Gelman et al. papers, arguing only what's wrong with the paper.

For the paper 'Sure-Thing Doubts' explain one argument that rational preferences must conform to the Independence Axiom, and McClennen's response to that argument. How could the proponent of the Independence Axiom respond?

3/2 Session 5: Learning preferences - Prof. Russell and Prof. Kariv

- More detail on how the standard model is implemented in AI systems: logical agents, probabilistic agents, learning agents
- The structure of preferences:
  - O multiattribute utility theory
  - O goals and ceteris paribus preferences
  - O utility over time
- A brief mention of Kahneman's psychological experiments on utility over time
- Finding out what preferences an agent has:
  - O Preference elicitation
  - O Inverse reinforcement learning

AIMA, Chapter 16, 17 (including Multi-Attribute Utility Theory)
Additional background and perspectives on AI (particularly for those who have not taken an AI course):
-
HC, Appendices A, B, C, D.

Suggested Readings:

Mike Wellman and Jon Doyle, Preferential Semantics for Goals, AAAI 91.
Georg von Wright, “The logic of preference reconsidered,” *Theory and Decision* 3 (1972): 140–67

Boutilier et al "CP-nets: A tool for representing and reasoning with conditional **ceteris paribus preference** statements", JAIR, 21, 2004.

John Harsanyi: “Games with incomplete information played by ‘Bayesian’ players, Parts I–III,” *Management Science* 14 (1967, 1968): 159–82, 320–34, 486–502.

Harsanyi, John C. "Welfare economics of variable tastes." *The Review of Economic Studies* 21.3 (1953): 204-213.

Richard Cyert and Morris de Groot, “Adaptive utility,” in *Expected Utility Hypotheses and the Allais Paradox*, ed. Maurice Allais and Ole Hagen (D. Reidel, 1979).

3/9 Session 6: Changing Preferences, Preferences about Preferences, and Intertemporal consistency - Prof. Kariv and Prof. Buchak

Richard Pettigrew, Choosing for Changing Selves, chs. 1, 3-6

Homework question:
Give an example of preference change over time (a new one, not one mentioned in Pettigrew’s book), and explain what one of the solutions Pettigrew discusses (Unchanging Utility, Utility of Utility, One True Utility, Aggregating Preferences, Aggregating Evaluation Functions, Aggregating Credences and Utilities) would say about the example.

3/16 Session 7: Other-regarding preferences - Prof. Russell and Prof. Kariv

HC Chapter 9 pages 215-216 (loyal AI) and 227-231 (nice, nasty, envious humans),
and students might like to skim the material on p211-214 and 217-227 (utilitarian AI and challenges to utilitarianism)
-a lot of Shachar papers

Fisman, Raymond, Shachar Kariv, and Daniel Markovits. "Individual preferences for giving." *American Economic Review* 97.5 (2007): 1858-1876.

Fisman, Raymond, Pamela Jakiela, and Shachar Kariv. "Distributional preferences and political behavior." *Journal of Public Economics* 155 (2017): 1-10.

Fisman, Raymond, Pamela Jakiela, and Shachar Kariv. *The distributional preferences of Americans*. No. w20145. National Bureau of Economic Research, 2014.

Suggested Readings:

Hori, Hajime. "Nonpaternalistic altruism and functional interdependence of social preferences." *Social Choice and Welfare* 32.1 (2009): 59-77.

Fisman, Raymond, et al. "The distributional preferences of an elite." *Science* 349.6254 (2015): aab0096.

Li, Jing, William H. Dow, and Shachar Kariv. "Social preferences of future physicians." *Proceedings of the National Academy of Sciences* 114.48 (2017): E10291-E10300.

Fisman, Raymond, Pamela Jakiela, and Shachar Kariv. "How did distributional preferences change during the great recession?." *Journal of Public Economics* 128 (2015): 84-95.

Short Memo #1 Due MONDAY MARCH 30

3. How do we model what multiple humans want?

3/30 Session 8: Game Theory - Prof. Kariv and Prof. Russell

Shachar lecture topics:

- Tragedy of the commons
- Keynes beauty contest
- Oligopolistic competition
- Auctions

Brandenburger, Adam M., and Barry J. Nalebuff. *The right game: Use game theory to shape strategy*. Vol. 76. Chicago: Harvard Business Review, 1995.

Stuart lecture topics:
1) motivation for why game theory is different from normal single-agent utility-maximization (soccer example);
Nash equilibrium including solving the soccer example; explain Prisoners' Dilemma, mention Tragedy of the Commons (HC p27-32)
2) IRL briefly, then assistance games, including paperclip game, off-switch game (HC p190-200)

Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell, Cooperative Inverse Reinforcement Learning. In *Advances in Neural Information Processing Systems 25*, MIT Press, 2017.
Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell, ``The off-switch game." In *Proc. IJCAI-17*, Melbourne, 2017.

Suggested Readings:
"Models in Microeconomic Theory" (a textbook for advanced undergraduates) by Martin Osborne and Ariel Rubinstein was published yesterday. The book can be downloaded for free from the Press' site. The book has two versions: in the "he" language and in the "she" language

4/6 Session 9: Social Choice Theory - Prof. Holliday

Sen, Amartya. "The possibility of social choice." *American Economic Review* 89.3 (1999): 349-378.

-students should review Wes's first set of slides (on bCourses -> Files -> Lecture Notes), since we'll be working with the same mathematical framework

Slides for this lecture based on Social Choice Theory.pdf

Homework question:
In his Nobel Prize lecture, Sen says: "Bentham's concern—and that of utilitarianism in general—was with the *total utility* of a community. This was irrespective of the distribution of that total, and in this there is an informational limitation of considerable ethical and political importance. For example, a person who is unlucky enough to have a uniformly lower capability to generate enjoyment and utility out of income (say, because of a handicap) would also be given, in the utilitarian ideal world, a *lower* share of a given total. This is a consequence of the single-minded pursuit of maximizing the sum-total of utilities (on the peculiar consequences of this unifocal priority, see Sen, 1970a, 1973a; John Rawls, 1971; Claude d'Aspremont and Louis Gevers, 1977)." Do you agree with Sen that utilitarianism involves an "informational limitation of considerable ethical and political importance"? Explain your answer.

4/13 Session 10: Interpersonal Comparisons and Welfarism - Prof. Holliday

Sen, "The Impossibility of a Paretian Liberal"

Slides for this lecture based on Social Choice Theory sessions 1 and 2.pdf

Homework question:
Sen takes the moral of his impossibility theorem to be that "in a very basic sense liberal values conflict with the Pareto principle" (p. 157). Do you agree with Sen's assessment? Explain your answer.

4/20 Session 11: Inequality - Prof. Buchak

Required:
Harsanyi, "Cardinal Utility in Welfare Economics and the Theory of Risk-Taking"
Rawls, *A Theory of Justice*, pp 52-56, 118-139
Buchak, "Taking Risks Behind the Veil of Ignorance"

Handout:
AGGREGATION HANDOUT (under Section 3 Materials on BCourses)

Optional:

Harsanyi, “Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls’s Theory”
Rawls, “Some Reasons for the Maximin Criteria”
Parfit, “Equality and Priority”

Homework question:
Which distributive rule do you think should be chosen behind a ‘veil of ignorance’ and why?

4. Wrap up

4/27 Session 12: All professors wrap up

Homework question:
Reflect on the single most important idea you have learned, as well as identify a topic that you wish had been covered but was not discussed (make sure to defend your choice!). Finally, write down at least one question you would like to discuss on the final day--we will go around and discuss them with all four professors taking part.

Short Memo #2 Due FRIDAY MAY 1

5/11 2-5pm Presentations by students of work-in-progress final papers (20 minutes presentation + 10 minutes questions from professors/other students)

Final Paper Due MONDAY MAY 18