2025 Fall Individual Report FAS-COMPSCI 2881R-Topics in Foundations of ML: AI Alignment and Safety 001 Boaz Barak

Project Title: **2025 Fall Harvard FAS Course Evaluation**

Course Audience: **65**
Responses Received: **44**
Response Ratio: **68 %**

Report Comments

Note:
The order that the questions appear on this report is not the same as the way the questions were displayed to students.
The order has been changed to make the report more readable.

Creation Date: **Wednesday, December 24, 2025**

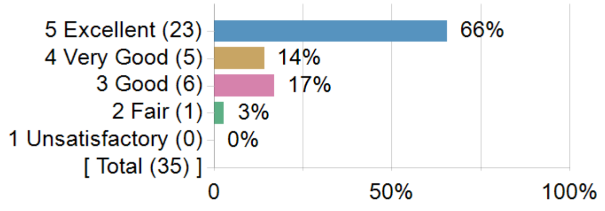
General Course Questions

Course General Questions

	Count	Excellent	Very Good	Good	Fair	Unsatisfactory	Course Mean	Dept Mean	Division Mean
Evaluate the course overall.	35	66%	14%	17%	3%	0%	4.43	3.99	3.99
Course materials (readings, audio-visual materials, textbooks, lab manuals, website, etc.)	34	62%	21%	15%	3%	0%	4.41	4.13	4.03
Assignments (exams, essays, problem sets, language homework, etc.)	32	47%	28%	22%	3%	0%	4.19	3.79	3.79
Feedback you received on work you produced in this course	32	38%	16%	19%	25%	3%	3.59	3.68	3.67
Section component of the course	12	75%	8%	8%	8%	0%	4.50	4.02	3.95

Evaluate the course overall.

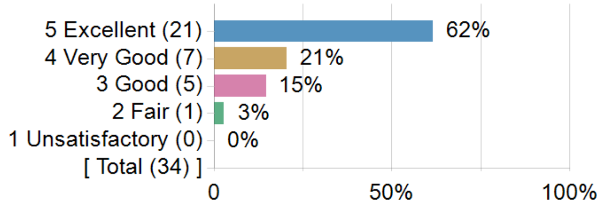
Evaluate the course overall.



Options	Score	Count	Percentage
Excellent	5	23	66%
Very Good	4	5	14%
Good	3	6	17%
Fair	2	1	3%
Unsatisfactory	1	0	0%
Statistics			**Value**
Response Ratio			54%
Mean			4.43
Median			5.00
Standard Deviation			0.88

Course materials (readings, audio-visual materials, textbooks, lab manuals, website, etc.)

Course materials (readings, audio-visual materials, textbooks, lab manuals, website, etc.)



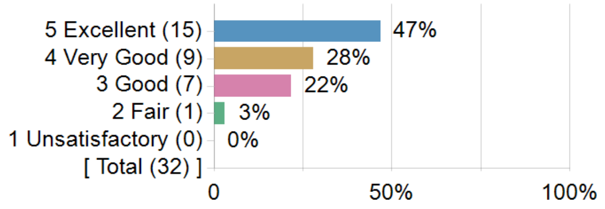
Options	Score	Count	Percentage
Excellent	5	21	62%
Very Good	4	7	21%
Good	3	5	15%
Fair	2	1	3%
Unsatisfactory	1	0	0%
Statistics			Value
Response Ratio			52%
Mean			4.41
Median			5.00
Standard Deviation			0.86

Add comments about course materials.

Comment
Perusall was great.
I liked Perusall for discussions! However, I like reading things on a tablet or via printing out, so that was kind of counterproductive. Maybe we could have a better way for discussions?
Readings were great. Perusell was good.
I'd strongly suggest excerpting papers. Having 150+ pages of reading means that people start skimming early.
I learned so much from the readings and really enjoyed engaging with others through Perusall.

Assignments (exams, essays, problem sets, language homework, etc.)

Assignments (exams, essays, problem sets, language homework, etc.)



Options	Score	Count	Percentage
Excellent	5	15	47%
Very Good	4	9	28%
Good	3	7	22%
Fair	2	1	3%
Unsatisfactory	1	0	0%

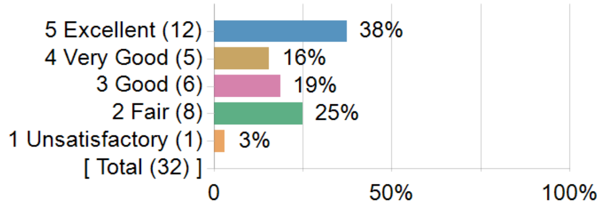
Statistics	Value
Response Ratio	49%
Mean	4.19
Median	4.00
Standard Deviation	0.90

Add comments about course assignments.

Comment
they felt a bit disjoint from what happened in the class. I think the range of possible projects should have been (even) higher, and the final project instructions should have been announced much earlier, so that students have more time for the final project.
This course would benefit with more pset–like assignments (i.e., the mini–project and final project were so much fun)
I love the format of the mini–project and the final project. Only thing is that it would've been great to get slightly more time to work on our final project.
I wish there were more assignments though, instead of just the two projects.

Feedback you received on work you produced in this course

Feedback you received on work you produced in this course



Options	Score	Count	Percentage
Excellent	5	12	38%
Very Good	4	5	16%
Good	3	6	19%
Fair	2	8	25%
Unsatisfactory	1	1	3%

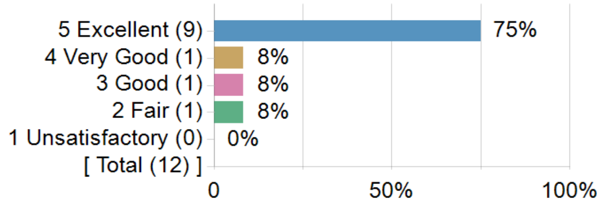
Statistics	Value
Response Ratio	49%
Mean	3.59
Median	4.00
Standard Deviation	1.32

Add comments about course feedback.

Comment
Barely any feedback
Feedback was thorough and helpful.
TF's could have bee more present (except Roy – he did a great job).
We never really got detailed feedback about the mini projects or the final projects. It was also hard to know how you were doing with respect to peruall expectations.
There was not a tremendous amount of feedback, but the feedback given was specific and helpful.
Appreciated the in–depth comments on course projects.

Section component of the course

Section component of the course



Options	Score	Count	Percentage
Excellent	5	9	75%
Very Good	4	1	8%
Good	3	1	8%
Fair	2	1	8%
Unsatisfactory	1	0	0%

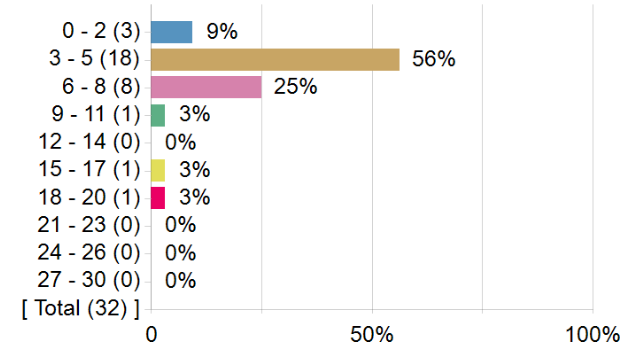
Statistics	Value
Response Ratio	18%
Mean	4.50
Median	5.00
Standard Deviation	1.00

Requirements - What did this course require of you?

On average, how many hours per week did you spend on coursework outside of class? Enter a whole number between 0 and 168.

Frequency chart and mean excludes students who answered 31 or more hours.

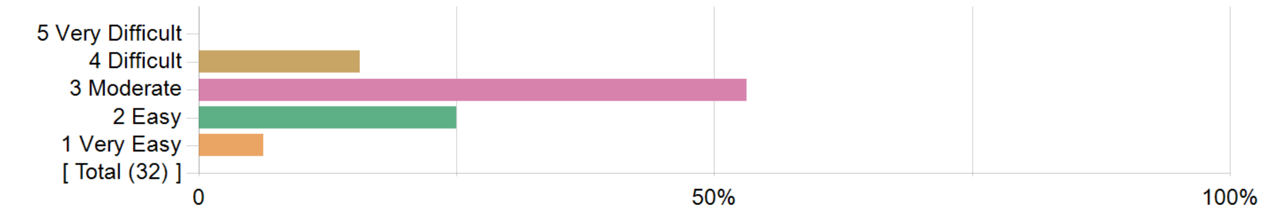
On average, how many hours per week did you spend on coursework outside of class? Enter a whole number between 0 and 168.



Statistics	Value
Response Count	32
Response Ratio	49%
Mean	5.47
Median	5.00
Mode	5
Standard Deviation	3.65

How difficult did you find this course?

How difficult did you find this course?



Options	Score	Count	Percentage
Very Difficult	5	0	0%
Difficult	4	5	16%
Moderate	3	17	53%
Easy	2	8	25%
Very Easy	1	2	6%

Statistics	Value
Response Ratio	49%
Mean	2.78
Median	3.00
Standard Deviation	0.79

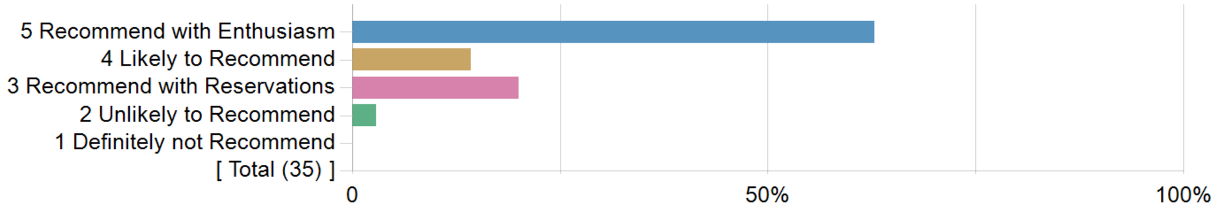
What was/were your reason(s) for enrolling in this course? (Please check all that apply)

Options	Count
Elective	22
Concentration or Department Requirement	14
Secondary Field or Language Citation Requirement	0
Undergraduate General Education Requirement	0
Expository Writing Requirement	0
Foreign Language Requirement	0
Pre-Med Requirement	0
Divisional Distribution Requirement	0
Quantitative Reasoning with Data Requirement	0

Recommendations - Would you recommend this course?

How strongly would you recommend this course to your peers?

How strongly would you recommend this course to your peers?



Options	Score	Count	Percentage
Recommend with Enthusiasm	5	22	63%
Likely to Recommend	4	5	14%
Recommend with Reservations	3	7	20%
Unlikely to Recommend	2	1	3%
Definitely not Recommend	1	0	0%

Statistics	Value
Response Ratio	54%
Mean	4.37
Median	5.00
Standard Deviation	0.91

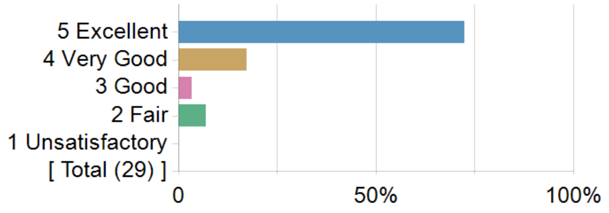
Evaluation of Instructors

General Instructor Questions

			Very				Instructor	Dept	Division
	Count	Excellent	Good	Good	Fair	Unsatisfactory	Mean	Mean	Mean
Evaluate your Instructor overall.	29	72%	17%	3%	7%	0%	4.55	4.40	4.34
Gives effective lectures or presentations, if applicable	27	70%	11%	11%	7%	0%	4.44	4.30	4.24
Is accessible outside of class (including after class, office hours, e-mail, etc.)	24	54%	13%	21%	8%	4%	4.04	4.27	4.29
Generates enthusiasm for the subject matter	28	82%	11%	7%	0%	0%	4.75	4.51	4.47
Facilitates discussion and encourages participation	27	74%	19%	4%	0%	4%	4.59	4.49	4.43
Gives useful feedback on assignments	16	56%	6%	19%	13%	6%	3.94	4.37	4.28
Returns assignments in a timely fashion	16	69%	19%	13%	0%	0%	4.56	4.30	4.17

Instructor

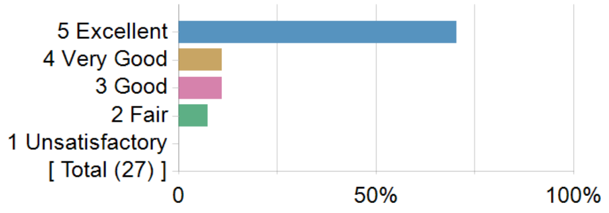
1. Evaluate your Instructor overall.



Options	Score	Count	Percentage
Excellent	5	21	72%
Very Good	4	5	17%
Good	3	1	3%
Fair	2	2	7%
Unsatisfactory	1	0	0%

Statistics	Value
Response Ratio	45%
Mean	4.55
Median	5.00
Standard Deviation	0.87

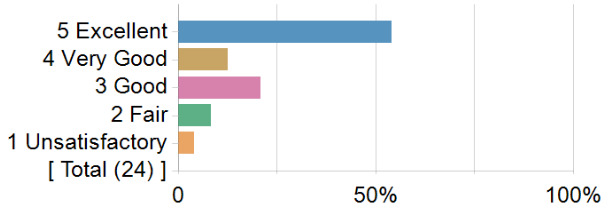
2. Gives effective lectures or presentations, if applicable



Options	Score	Count	Percentage
Excellent	5	19	70%
Very Good	4	3	11%
Good	3	3	11%
Fair	2	2	7%
Unsatisfactory	1	0	0%

Statistics	Value
Response Ratio	42%
Mean	4.44
Median	5.00
Standard Deviation	0.97

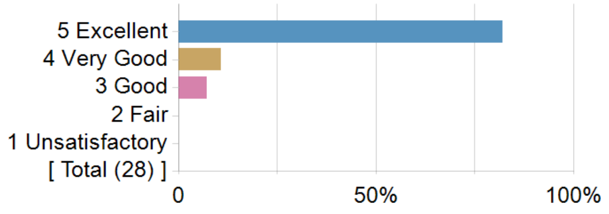
3. Is accessible outside of class (including after class, office hours, e-mail, etc.)



Options	Score	Count	Percentage
Excellent	5	13	54%
Very Good	4	3	13%
Good	3	5	21%
Fair	2	2	8%
Unsatisfactory	1	1	4%

Statistics	Value
Response Ratio	37%
Mean	4.04
Median	5.00
Standard Deviation	1.23

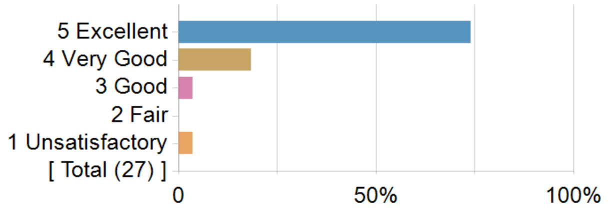
4. Generates enthusiasm for the subject matter



Options	Score	Count	Percentage
Excellent	5	23	82%
Very Good	4	3	11%
Good	3	2	7%
Fair	2	0	0%
Unsatisfactory	1	0	0%

Statistics	Value
Response Ratio	43%
Mean	4.75
Median	5.00
Standard Deviation	0.59

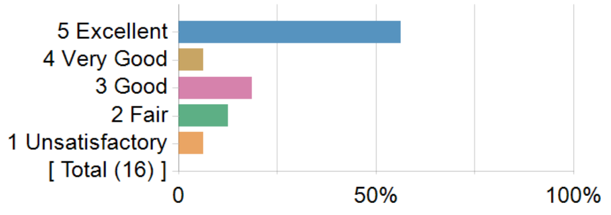
5. Facilitates discussion and encourages participation



Options	Score	Count	Percentage
Excellent	5	20	74%
Very Good	4	5	19%
Good	3	1	4%
Fair	2	0	0%
Unsatisfactory	1	1	4%

Statistics	Value
Response Ratio	42%
Mean	4.59
Median	5.00
Standard Deviation	0.89

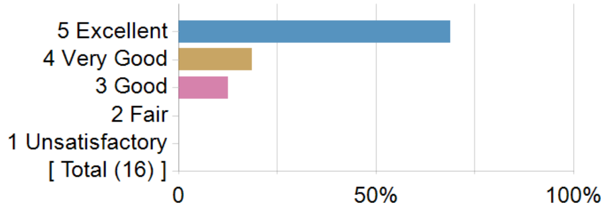
6. Gives useful feedback on assignments



Options	Score	Count	Percentage
Excellent	5	9	56%
Very Good	4	1	6%
Good	3	3	19%
Fair	2	2	13%
Unsatisfactory	1	1	6%

Statistics	Value
Response Ratio	25%
Mean	3.94
Median	5.00
Standard Deviation	1.39

7. Returns assignments in a timely fashion



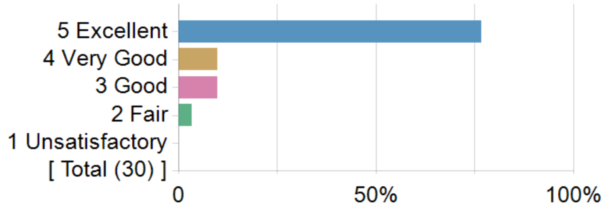
Options	Score	Count	Percentage
Excellent	5	11	69%
Very Good	4	3	19%
Good	3	2	13%
Fair	2	0	0%
Unsatisfactory	1	0	0%

Statistics	Value
Response Ratio	25%
Mean	4.56
Median	5.00
Standard Deviation	0.73

GSAS Module Questions

Included discussion or assignments that pointed to a potential dissertation topic, or, in the sciences, a potential research lab

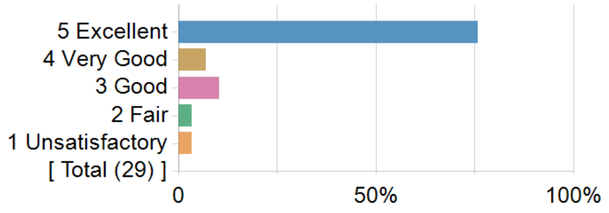
Included discussion or assignments that pointed to a potential dissertation topic, or, in the sciences, a potential research lab



Options	Score	Count	Percentage
Excellent	5	23	77%
Very Good	4	3	10%
Good	3	3	10%
Fair	2	1	3%
Unsatisfactory	1	0	0%
Statistics			Value
Mean			4.60
Median			5.00
Standard Deviation			0.81

Included assignments that helped to develop necessary research skills for a potential dissertation topic

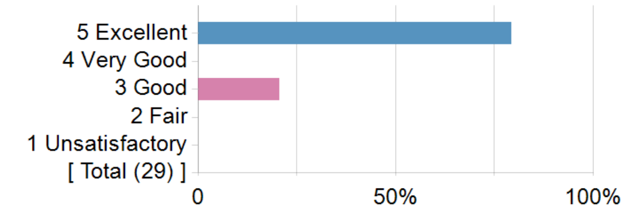
Included assignments that helped to develop necessary research skills for a potential dissertation topic



Options	Score	Count	Percentage
Excellent	5	22	76%
Very Good	4	2	7%
Good	3	3	10%
Fair	2	1	3%
Unsatisfactory	1	1	3%
Statistics			Value
Mean			4.48
Median			5.00
Standard Deviation			1.06

Included assignments that could be developed into professional talks or potential publications

Included assignments that could be developed into professional talks or potential publications

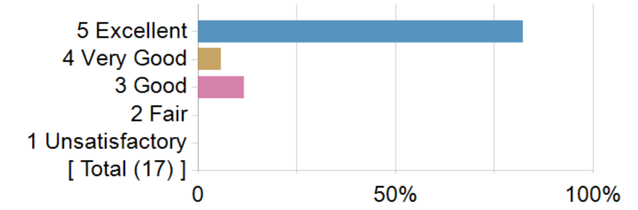


Options	Score	Count	Percentage
Excellent	5	23	79%
Very Good	4	0	0%
Good	3	6	21%
Fair	2	0	0%
Unsatisfactory	1	0	0%

Statistics	Value
Mean	4.59
Median	5.00
Standard Deviation	0.82

Helped directly or indirectly in preparation for generals

Helped directly or indirectly in preparation for generals



Options	Score	Count	Percentage
Excellent	5	14	82%
Very Good	4	1	6%
Good	3	2	12%
Fair	2	0	0%
Unsatisfactory	1	0	0%

Statistics	Value
Mean	4.71
Median	5.00
Standard Deviation	0.69

Comment on aspects of the course as they relate to professional development, including preparation for future teaching.

Comment
Really appreciated the focus on developing longterm research projects.

General Course Questions - Comments

What were the strengths of this course? Please be specific and use concrete examples where possible.

Comment
Good overview of AI safety, technical without being overbearing
This helped me understand how the tech industry and adjacent labs currently think about AI safety. This is useful and really cool. I also appreciate how the class is experimental and picking up current debates.
Boaz was great, and I loved the project–focused approach of the course, having assignments all be small experiments.
– great lecturers – fascinating and relevant topic area – thoughtfully selected topics and reading
All of the guest speakers gave very interesting presentations and it was great to hear from alignment researchers in the AI safety community as well as scholars from other disciplines (I found the economists' presentations and Ziad's presentation particularly eye–opening). I also enjoyed the pre–readings–– since there is a lot of research activity, I found it helpful for the teaching staff to direct me to the most important readings within AI safety.
The course gives a comprehensive overview of the salient topics in AI safety and alignment. The guest lectures were definitely a highlight for me. We have had many high profile speakers including Neel Nanda, Jack Lindsay, Nicolas Carlini, and many others. The pre–lecture readings were also very informative and have taught me a lot about all the different topics.
Very good selection of papers for discussion, incredible speakers. Loved the candid perspective from the Prof. Barak.
Guest speakers
Learned a great deal about not just the field of AI safety but also methods, terms, and practices of the field of applied AI research. Loved the readings, informal style of instruction, and the projects.
Very technical, great speakers, great selection of students and amazing teaching staff.
The course was an excellent survey of recent research in AI safety and the current state of the field. The guest lectures were great, and the seminar style worked well.
Presents the topic of AI safety from many different angles
Really interesting material, Prof. Barak is clearly passionate about the topic, and we had such interesting readings and discussions and best of all, very interesting speakers come to talk about their work and research.
interesting content; boaz is a legend and he assembled the avengers of guest speakers; also the workload is manageable
Professor Boaz was a knowledgeable and engaging lecturer and the guest speakers were all fantastic. Readings were informative and not too overwhelming for someone who was just getting started in the AI safety sphere.
We heard from some amazing speakers in the field, and Boaz's technical lectures were really excellent.

How could this course be improved? Please use concrete examples where possible and provide constructive suggestions.

Comment
I thought it was a well done graduate seminar. Nothing insane, but very well done.
This course's main problem is that it takes current industry work on AI safety as gospel. Since so much of this work is highly speculative and riddled with problems, this course should have been much more geared towards controversial, in–class, whole–cohort discussion. We had very few of those. – Almost EVERY course meeting was RECORDED and PUT ON YOUTUBE. That is one way to make sure discussions will never become too controversial. – This should have been taught in a room that allows for whole–cohort discussions. The group tables in LL2.229 are not good for this, as it's hard to see the rest of the cohort. – There should have been much more time for discussions, both in–class with the cohort and with the guest speakers. It should be much more encouraged to question the work that guest speakers have done. As far as assignments go, they felt a bit disjoint from what happened in the class. I think the range of possible projects should have been (even) higher, and the final project instructions should have been announced much earlier, so that students have more time for the final project.
I think lectures could use being a bit more technical, such as interpretability techniques etc.
– more upfront organizational clarity :)
Readings were assigned late in the week, assignment structure wasn't decided until very late in the semester.
I thought it would have been nice for the grading policy and expectations particularly for the mini–project to be made more clear in advance. For the mini–project, our group wasn't quite sure how closely the reproduction of results needed to be and ended up seeking clarification from the teaching staff closer to the deadline. I recognize that this was the first iteration of the course so some of the grading details were ironed out as the semester progressed, but for future classes I think it would be helpful to provide grade breakdowns in the syllabus and a clearer rubric/understanding of what the teaching staff is looking for in both projects.
I want to start by noting that it is already an amazing course as it is. The only thing is that maybe we would get slightly longer to work on our final project or at least receive information on it a bit earlier.
The course was a bit chaotic. Communication was via Slack, readings on perusal, assignment submissions on google forms, syllabus on another personal website, and nothing at all on Canvas. Centralised use of one platform would have been ideal. The projects were not on the syllabus at the start of the semester, which means that they came as a bit of a surprise. It was unfortunate, as this meant that I had another team project to do that I had not budgeted for.
Having the projects start earlier allows students to work on them for longer and achieve better results.
The class's evaluations were very ad hoc and in many cases had changes to the requirements or deadlines announced with less than a week to go. This is largely because the course's assignments were devised by one (great, hardworking) TF, while the other TFs and the professor were rather tuned–out.
The course could have been more technical and deeper.
i thought i would have benefited from the technical content of week 2 being expanded upon
The class could have offered more opportunities for students to present / run experiments / write blogs.
I thought that it would have been good to have had more opportunities to build technical skills relevant for AI safety, possibly in the form of problem sets or more technical lectures. At times, the lectures felt very speculative, especially those on societal impacts.

Requirements Comments - What did this course require of you?

In your opinion, what preparation or background is necessary to take this course?

Comment
Coding, understanding of research/graduate maturity
Experience in practical AI work
A good background on actually building things!
I think this course can be taken with background in a wide variety of disciplines, but moderate Python programming experience and experience reading and understanding academic papers is required.
Some prior research experience!
It's good to have a general understanding of how LLMs work and the relevant training approaches such as RL and SFT etc. However, I think the pre–lecture is very well chosen such that if you put in the work and do the reading it will equip you with the relevant knowledge anyways.
Basic AI and ML
Anything? It tried to be interdisciplinary and doesn't use any real math
Solid foundation in ML/AI fundamentals.
Machine Learning, Previous experience with ML Research, Reinforcement Learning, etc.
Prior experience with ML is extremely helpful but probably not strictly required.
strong programming skills and math ability to truly understand the readings
Substantial Python coding experience and familiarity with using open source AI models. Familiarity with conducting research, and some exposure to AI research. Readiness to read academic papers.
Reading papers
Ability to fine–tune models and some idea of how to run experiments with them.

Recommendations Comments - Would you recommend this course?

What did you take away from your experience in this course? What did you learn? How did this course change you?

Comment
I learned a lot about AI safety
I found learning about and gaining practice with different frameworks of thinking within AI safety and alignment research (e.g. model specs, policies, ablation studies) along with learning about cutting–edge research directions in the field were extremely interesting. Entering the course, I had been interested in the potential social impacts of AI deployment like algorithmic bias, environmental impact, and data privacy, so learning about AI safety from a more technical perspective was eye–opening. The course helped me clarify the relationship between the technical research and these more practical social or policy concerns.
I have gained a good understanding of the frontier topics related to AI safety and alignment. I was exposed to a range of different views of AI development and different directions in which people can work on to make AI safer. It has further strengthened my desire to work in AI safety research.
I learned so much about the current research in ML and AI alignment, discovered a new research direction, got to interact with speakers from different frontier and research labs.
i learned a lot about the field of AI alignment/safety, some new concepts technically, and how to produce sound AI safety research
I learned not to fear the rapid growth of AI but rather approach it with a healthy level of cautious optimism. I think different people had different takeaways from this class, but I personally came out feeling invigorated with a more realistic understanding of how AI's progression might impact society. I think the greatest fear and responsibility lies, as always, with the users of this tool and the humans who could wield AI for so much good and also potentially do so much harm. I am excited to live in this moment and be a part of a generation that can shape the capabilities of AI.

What would you like to tell future students about this class? (Your response to this question may be published anonymously.)

Comment
Boaz is a great professor. I really appreciate all that they did to make this a quality course – from lining up quality guest lecturers, to providing OpenAI credits and compute reimbursements, to 'little' (but very, very much appreciated) things like catering great snacks and coffee for each lecture. Boaz clearly cares. This was a well done graduate seminar. As someone who isn't super crazy about safety as a line of research / not extremely scared about risks stemming from AI, I can't say that this course inspired me to entirely shift my research interests, but I appreciated a very high quality introduction to a variety of perspectives in this line of research and

approaches in AI safety and alignment.

The readings were well chosen, and again I really appreciated the quality of the guest lecturers and Boaz's interactions with them, I think they were highly informative. The readings were required on Panopto, but I wouldn't be too concerned about that. The mini projects were a fun way to get hands on, and the final project expectations were reasonable and well communicated. I have no idea what the grading for the course is, but for concurrent master's / graduate students I wouldn't think that's a reason not to take the course.

Overall if you have any interest in AI safety or foundation models in general, I would highly recommend!

Very fun class, great intro to AI Safety and plenty of bones thrown at people who are already more familiar with the field. Both Boaz's lectures and the (high–profile!) guest lectures were very informative, and both the mini–project and the final project were fun to do, as well as the weekly student experiments.

I honestly wish there was a bit more work: more mini–projects would help all of us get more hands–on exposure to the concepts and tease apart new interesting research directions in safety. Excited for this class to come back next year, more people should take this and research AI safety!

This was a great introduction to AI safety. Every week Boaz picked some very up to date papers, including new ones published that same week. We surveyed a broad range of topics from economic impacts of AI to mech interp and adversarial attacks. There was a mini–project due around Thanksgiving, a final project due around reading period, and lesson experiments one can sign up for; these are all very informative and a wonderful way to get to know your classmates. The lectures themselves were sometimes hit or miss depending on the guest lecturer and your personal interest; and I wished the class offered more opportunities for students to socialize and interact with one another (for example have a poster session for the mini–project as well). Overall very eye–opening and it is a good effort, enjoy it if you get in!

This is a good course to take if you are interested in how the tech industry and adjacent labs think about AI safety.

It is a super experimental course with a syllabus that was constantly updated. I liked that.

If you want to learn about AI safety and discuss so you can come up with your own opinion, you may find this class a bit frustrating.

This course's main problem is that it takes current industry work on AI safety as gospel. Since so much of this work is highly speculative and riddled with problems, this course should have been much more geared towards controversial, in–class, whole–cohort discussion. We had very few of those.
– Almost EVERY course meeting was RECORDED and PUT ON YOUTUBE. That is one way to make sure discussions will never become too controversial.
– This should have been taught in a room that allows for whole–cohort discussions. The group tables in LL2.229 are not good for this, as it's hard to see the rest of the cohort.
– There should have been much more time for discussions, both in–class with the cohort and with the guest speakers. It should be much more encouraged to question the work that guest speakers have done.

As far as assignments go, they felt a bit disjoint from what happened in the class. I think the range of possible projects should have been (even) higher, and the final project instructions should have been announced much earlier, so that students have more time for the final project.

This class really should have been much better at encouraging discussions and open–mindedness.

This class would benefit from psets or more mini–projects. The HW 0 was great and a lot of fun, but I——along with other classmates——were surprised to see that the actual class ended up being completely reading–based. The lectures also felt too long at times. But the mini–project and project were amazing.

Great class, with a great instructor — Boaz makes every single lesson a joy to look forward to, and the industry lecturers are something you can only get here at Harvard: we had a ton of just fascinating guest lectures!

Great class. Boaz is very knowledgable and brings in great guest speakers. He enjoys facilitating thoughtful discussion and puts most of the emphasis on that. Readings were thoughtful, actively curated, and very current. Projects and grading were a little more disorganized so maybe not the best class to take for structure and clarity on that front, but well worth it for the course content.

AMAZING CLASS. Great survey of AI safety, learned a lot. Boaz is very well connected, got a bunch of AI safety big–shots to give guest lectures. Super amazing course. Quite well organized considering it's the first year it's offered.

Take CS 2881! It was a great course and I learned a lot from Prof. Barak and all the guest speakers (from OpenAI, Anthropic, DeepMind, etc). The 3 hour class is tough, but Boaz is an engaging lecturer and is not only thoughtful about how he conceptualizes different aspects of AI safety (which at the moment are pretty broad and open to discussion/interpretation) but also values the questions, opinions, and work of everyone in the class. I would suggest stopping by his office hours too to ask Boaz questions and to have interesting discussions— they were always very relaxed.

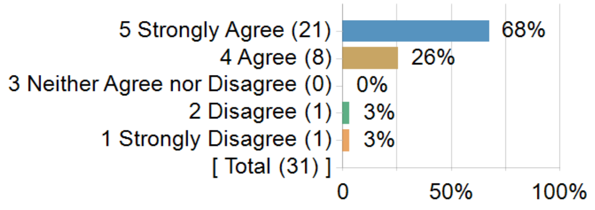
I came into the class as a CS undergrad (senior), and this was my first CS grad research class— it was really valuable for me to get some hands–on experience running experiments like an AI alignment researcher would and I think a lot of folks in the class were able to clarify their research interests. I wouldn't be afraid of taking this class if you don't think you have enough preparation— just some Python experience is sufficient since AI usage is generally encouraged in the course.

Workload was very manageable (~1–2 hrs/wk most weeks, but 10–15 hrs total for projects/experiments) but could change since it is a newer class–– in this iteration there were two projects (a mini–project during midterms and a final project) along with an optional student experiment where you can sign up to give a 20–minute group presentation empirically testing a research question related to the topic for that week's lecture (I'd recommend doing this too, it was fun!).
I have not recommended a class with such enthusiasm in a very long time. If there is one class at Harvard you should take it is this one. Boaz is absolutely amazing. He is incredibly knowledgeable in this field (and many other fields) and is just a great person in general. The guest lectures were also really cool, with a line up of high profile speakers such as Neel Nanda, Jack Lindsay and Nicolas Carlini. If you're hesitating about taking this class, my recommendation would be to just do it. It's really amazing and the pre–lecture readings were very well chosen such that if you spend the time doing the readings you'll be well–equipped with the necessary background to learn about the topics each week.
this is a great class and both Boaz + the guest lecturers are amazing; class is a bit unorganized at times but its still very manageable
this is a pretty awesome class, though idk if itll ever be taught again boaz works at openai, so every lecture (meets once a week) was about a cool topic in ai safety (the website is public online, so you can read/watch through the past lecture topics) lectures often were by guest speakers from openai/anthropic/redwood research/metr, etc – not every class you get an opportunity like that workload was pretty chill. just a mini project and final project – you make of it what yo uwant. optional experiment to present to the class too overall, this class inspired me a ton about what the future of ai might look like. gives you more of a sense of the conversations that are happening in the rooms at openai/anthropic/etc
Fake, but in an interesting way. Don't expect to learn more ML, but expect to yap about it for a bit.
This is a fantastic course in which you will learn a ton! It is comparatively lighter lift than other grad courses, but you get out of it what you put in.
The state of the field is advancing very quickly, and this course was a good opportunity to read papers that otherwise I would not have had time to get through. It's also a good opportunity to test your ability to put together a novel research project on a short timeline. Needless to say, that short timeline also presents challenges. Feedback opportunities weren't extensive, and the class is worse because the students participating come from very diverse backgrounds, making it harder to find cohesive project groups.
chill class, Boaz is a legend. you get as much as you put in. there are a lot of readings that are deeply technical, you can choose to skim or read in depth, in which case you would gain a lot. lots of good speajers (NIck carlinini, neel nanda, etc). really the forefront of the field of ai safety.
Take this class!! You will be exposed to many perspectives and expert opinions regarding the trajectory of AI, what are reasonable and unreasonable fears, and what we are able to do in the process of technological growth. Professor Boaz is not only a fantastic teacher but also humorous and very approachable. The discussions and readings are informative and you will take from this class a clearer picture of where you stand. Possibly the most relevant and useful class I've taken!
Lots of AI Safety papers and guest speakers which was really cool! The content is still pretty speculative, but I'd imagine if this was being taught / repeated in the future for AI Safety topic, this will become increasingly relevant and news updates would be exciting.
This is a good class. It may be more organized in a subsequent year. I would expect to gain more technical skills in a technical elective—this course orients you somewhat to think about research problems, but a good deal of it is speculative (how will AI change the future etc. etc.)

Open Discourse in the Classroom

In this course, most students listen attentively with an open mind and a willingness to change their point of view as they learn more about the topic.

In this course, most students listen attentively with an open mind and a willingness to change their point of view as they learn more about the topic.



Options	Score	Count	Percentage
Strongly Agree	5	21	68%
Agree	4	8	26%
Neither Agree nor Disagree	3	0	0%
Disagree	2	1	3%
Strongly Disagree	1	1	3%

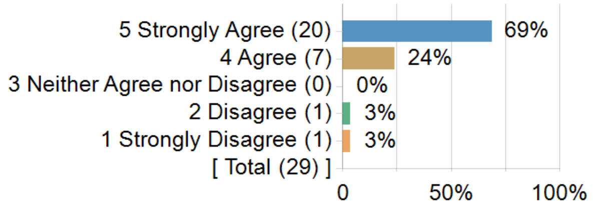
Statistics	Value
Response Ratio	48%
Mean	4.52
Median	5.00
Standard Deviation	0.93

Add comments about this question.

Comment
This class really should have been much better at encouraging discussions and open–mindedness. For example, Boaz kept talking about others who criticize climate impacts of LLMs (which he partially disagrees with) — why was there never a class discussion about this?
People were open about their opinions and very willing to listen and discuss with others. This was especially important for this class, where we were shaping our own views of how different actors should approach AI safety as the class went on, and I felt that the class environment fostered such thinking.

In this course (including sections), I feel comfortable expressing my views on controversial topics.

In this course (including sections), I feel comfortable expressing my views on controversial topics.



Options	Score	Count	Percentage
Strongly Agree	5	20	69%
Agree	4	7	24%
Neither Agree nor Disagree	3	0	0%
Disagree	2	1	3%
Strongly Disagree	1	1	3%
Statistics			Value
Response Ratio			45%
Mean			4.52
Median			5.00
Standard Deviation			0.95

Add comments about this question.

Comment
This class did nothing at all to encourage in–class discussion of its often highly controversial themes. Also, ALMOST EVERY COURSE MEETING WAS RECORDED AND PUT ON YOUTUBE.
People's willingness to share their personal views created a comfortable environment where controversial topics could be discussed with civility.

Instructor Comments

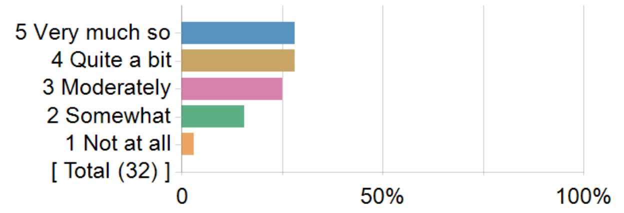
Please comment on this person's teaching. (Your response to this question may be published anonymously.)

Comment
Boaz teaches very well and creates a wonderful, collaborative learning environment in the classroom.
I would go as far as saying that Boaz is the best lecturer that I have ever had. He is so incredibly knowledgeable in AI safety and also just everything else in general. It is amazing to hear about his thoughts on the current progress in AI safety research and what directions he thinks are promising. He is also very funny and very effective at delivering the lectures.
boaz tough
Great. Very oganized and led the class very well.
Boaz is influential within the field, but honestly he seemed pretty checked–out from about halfway through the semester onwards. One of the TFs really did the major lifting for the course. Boaz's office hours were ad hoc and minimal, he often just let guest lecturers run the show, and he changed the topics of class and the readings assigned less than a week prior to the class meeting date.
He is very passionate and loves hearing people's opinions, and is wonderful to interact with!
boaz is a legend, very knowledgeable and full of insights on the field. so grateful to learn from him!
Boaz's enthusiasm and genuine interest in this subject shines throughout his teaching. I always looked forward to lecture, knowing that it would be the perfect balance of amusing, deeply reflective, and informative.

Custom Questions

Did the course impact your plans for your future studies or job?

Did the course impact your plans for your future studies or job?



Options	Score	Count	Percentage
Very much so	5	9	28%
Quite a bit	4	9	28%
Moderately	3	8	25%
Somewhat	2	5	16%
Not at all	1	1	3%

Statistics	Value
Mean	3.63
Median	4.00
Standard Deviation	1.16

Did this course change your views on AI or AI safety? Its importance, future, or likelihood to end up well? If so in what ways?

Comment
Yes. The course did convince me that alignment is closely tied with capabilities, so even for those of us focused on understanding capabilities it is important to be in touch with safety research since these approaches can be used to make better (as well as safer) models.
A lot! I think it's the world's most important problem now, and I want to work more with it in the near future.
it is important, might work in it.
This course made me realize the diversity of problems in AI safety that are yet to be solved. I really appreciated the scope of problems we discussed in class.
Yes! I learned a lot about AI safety and have new appreciation for just how difficult yet critical of a responsibility it is, particularly thinking about how we align models and to what. I also now feel that reading / studying or even performing AI safety research is more achievable than I ever would have thought previously, which is cool.
I have never subscribed to the idea that, at least in the short term, AI will rise up, take over, and kill us all. I have always seen the biggest risk associated with AI coming from us not understanding its behaviour properly, which can be seen from examples like emergent misalignment. This course has reaffirmed this idea and it is great to see Boaz holding a similar idea as I do.
Not really, but I gained the language to talk about it a little better.
Certainly. It made the field much more accessible, and showed me the methods that researchers are using. It is no longer a nebulous topic. The course also demonstrated that "AI safety" means a ton of different things. I think that AI safety is only going to increase in importance as AI becomes more embedded within societies, especially as these systems operate as black boxes and have non–deterministic outcomes.
I take some of the risks of misalignment more seriously, but on the whole my views didn't shift: AI, in the immediate future, is a normal technology, and the harms are likely to be in mundane but severe modalities: Job displacement, scams, cyberattacks, etc.
It gave me a better perspective. I don't think the course involved enough aspects of technical AI safety for me to develop a more serious understanding during the course.
yes. i of course think AI saftey is important, but now approach it from a different lense (before i was mech–interp pilled, now i realize there are other avenues to achieving safety besides hard core interp)
This course taught me to approach AI safety with a moderate mindset — not to demonize the technology nor blindly follow in its trajectory. I think AI has the potential to facilitate and simplify many jobs and the danger only begins when people start willingly forsaking their human agency to the decisions of an AI model. I think AI will only hold as much power as we humans are willing to give it, and possibly the most important decision for the future is exactly how much and what kind of power we should give it.
Not too much, still very speculative, but comfort that many people are thinking about it and perhaps taking small steps for preventative harm.

What are the questions or aspects of AI safety that you find most important?

Comment
How do we create models that understand human goals?
Interpretability and Model Spec design.
economic impacts, ai psycophancy/psychosis etc, interpretibility, adversarial attacks / alignment
How do we do scalable oversight / raise the bar on intelligence while keeping models safe? And, more promptly, how do we stop the negative effects from the models we already have started to observe, such as mass misinformation?
What happens after AI alignment? i.e. what can we do re: concentration of power/techno–feudalism?
I began most worried about AI in the hands of a malicious human, and I think I remain that way.
I am curious about how "LLM as judge" is going to evolve as a research technique, given that there are potential alignment issues arising from this as well – especially as models get larger and more capable. I am also wondering about the use of eval sets like GPQA, and whether one day there will need to be a set of evals that are kept in cold storage permanently offline to avoid any model from being trained against them.
how do we ensure that AIs grow to love human beings?
How to scale at a safe rate, balancing the development of better capabilities while not overreaching and entering unsafe territory? How to mitigate over–reliance on AI, particularly emotional reliance that could replace people's desire for human interaction to an unhealthy extent? How can we teach people to become safe users of AI — because safety is both the responsibility of the producer and the user?