The Greatest Good

David Gottlieb (dmg1@stanford.edu)

17 April 2024

Singer, “Famine, Affluence, and Morality”

Singer's strategy

- The strategy is to start from uncontroversial assumptions and reach radical conclusions.

Singer's strategy

- The strategy is to start from uncontroversial assumptions and reach radical conclusions.
- The uncontroversial assumptions:

Singer's strategy

- The strategy is to start from uncontroversial assumptions and reach radical conclusions.
- The uncontroversial assumptions:
  - Suffering and premature death are bad things

Singer's strategy

- The strategy is to start from uncontroversial assumptions and reach radical conclusions.
- The uncontroversial assumptions:
  - Suffering and premature death are bad things
  - If we can stop a bad thing without the sacrifice of anything of comparable moral import, we should.

Singer's strategy

- The strategy is to start from uncontroversial assumptions and reach radical conclusions.
- The uncontroversial assumptions:
  - Suffering and premature death are bad things
  - If we can stop a bad thing without the sacrifice of anything of comparable moral import, we should.
  - (Without weakening the conclusion, we can weaken the second principle to: if we can stop a bad thing without the sacrifice of anything of *significant* moral import, we should.)

Singer's strategy

- The strategy is to start from uncontroversial assumptions and reach radical conclusions.
- The uncontroversial assumptions:
  - Suffering and premature death are bad things
  - If we can stop a bad thing without the sacrifice of anything of comparable moral import, we should.
  - (Without weakening the conclusion, we can weaken the second principle to: if we can stop a bad thing without the sacrifice of anything of *significant* moral import, we should.)
- The drowning child example illustrates the second principle.

Singer's strategy

- The strategy is to start from uncontroversial assumptions and reach radical conclusions.
- The uncontroversial assumptions:
  - Suffering and premature death are bad things
  - If we can stop a bad thing without the sacrifice of anything of comparable moral import, we should.
  - (Without weakening the conclusion, we can weaken the second principle to: if we can stop a bad thing without the sacrifice of anything of *significant* moral import, we should.)
- The drowning child example illustrates the second principle.
- Note that none of the principles say anything like: if we can stop a bad thing *nearby* . . .

Singer's strategy

- The strategy is to start from uncontroversial assumptions and reach radical conclusions.
- The uncontroversial assumptions:
  - Suffering and premature death are bad things
  - If we can stop a bad thing without the sacrifice of anything of comparable moral import, we should.
  - (Without weakening the conclusion, we can weaken the second principle to: if we can stop a bad thing without the sacrifice of anything of *significant* moral import, we should.)
- The drowning child example illustrates the second principle.
- Note that none of the principles say anything like: if we can stop a bad thing *nearby* . . .
- Physical proximity matters *instrumentally* because sometimes we have a special opportunity to help someone close by.But it doesn't matter *per se*.

Singer's strategy

- The strategy is to start from uncontroversial assumptions and reach radical conclusions.
- The uncontroversial assumptions:
  - Suffering and premature death are bad things
  - If we can stop a bad thing without the sacrifice of anything of comparable moral import, we should.
  - (Without weakening the conclusion, we can weaken the second principle to: if we can stop a bad thing without the sacrifice of anything of *significant* moral import, we should.)
- The drowning child example illustrates the second principle.
- Note that none of the principles say anything like: if we can stop a bad thing *nearby* . . .
- Physical proximity matters *instrumentally* because sometimes we have a special opportunity to help someone close by.But it doesn't matter *per se*.
- As a person in a rich country in a globalized world, you can help people a lot even though you're not close to them at all.

Singer's strategy

- The strategy is to start from uncontroversial assumptions and reach radical conclusions.
- The uncontroversial assumptions:
  - Suffering and premature death are bad things
  - If we can stop a bad thing without the sacrifice of anything of comparable moral import, we should.
  - (Without weakening the conclusion, we can weaken the second principle to: if we can stop a bad thing without the sacrifice of anything of *significant* moral import, we should.)
- The drowning child example illustrates the second principle.
- Note that none of the principles say anything like: if we can stop a bad thing *nearby* . . .
- Physical proximity matters *instrumentally* because sometimes we have a special opportunity to help someone close by.But it doesn't matter *per se*.
- As a person in a rich country in a globalized world, you can help people a lot even though you're not close to them at all.
- Anyone who *can* help is *responsible* for helping.

Moral feelings and self-indulgence

Jim and the Indians

- Bernard Williams (1973): Jim is backpacking through Central America when he encounters a death squad led by Pedro that is getting ready to kill 20 Indian villagers. Pedro offers Jim a deal: you shoot one of the Indians and we'll spare the other 19. What should Jim do?

Jim and the Indians

- Bernard Williams (1973): Jim is backpacking through Central America when he encounters a death squad led by Pedro that is getting ready to kill 20 Indian villagers. Pedro offers Jim a deal: you shoot one of the Indians and we'll spare the other 19. What should Jim do?
- Does Jim do a wrong by shooting the 1 to spare the 19?

Jim and the Indians

- Bernard Williams (1973): Jim is backpacking through Central America when he encounters a death squad led by Pedro that is getting ready to kill 20 Indian villagers. Pedro offers Jim a deal: you shoot one of the Indians and we'll spare the other 19. What should Jim do?
- Does Jim do a wrong by shooting the 1 to spare the 19?
- If he feels bad about it, is he making a mistake?

The good doesn't care about your feelings

- Eliezer Yudkowsky (2008):

The good doesn't care about your feelings

- Eliezer Yudkowsky (2008):
  You know what? This isn't about your feelings. A human life, with all its joys and all its pains, adding up over the course of decades, is worth far more than your brain's feelings of comfort or discomfort with a plan. Does computing the expected utility feel too cold-blooded for your taste? Well, that feeling isn't even a feather in the scales, when a life is at stake. Just shut up and multiply...
  Altruism isn't the warm fuzzy feeling you get from being altruistic. If you're doing it for the spiritual benefit, that is nothing but selfishness. The primary thing is to help others, whatever the means. So shut up and multiply!

Get your fuzzies cheap

- Yudkowsky (2009):

Get your fuzzies cheap

- Yudkowsky (2009):
  There is this very, very old puzzle/observation in economics about the lawyer who spends an hour volunteering at the soup kitchen, instead of working an extra hour and donating the money to hire someone. . .
  If the lawyer needs to work an hour at the soup kitchen to keep himself motivated and remind himself why he's doing what he's doing, that's fine. But he should also be donating some of the hours he worked at the office, because that is the power of professional specialization and it is how grownups really get things done. One might consider the check as buying the right to volunteer at the soup kitchen, or validating the time spent at the soup kitchen.

Summary on fuzzies

- "Fuzzies" are our apparently moral feelings that don't follow a utilitarian calculus:

Summary on fuzzies

- "Fuzzies" are our apparently moral feelings that don't follow a utilitarian calculus:
  - Guilt at doing actions that are net-beneficial (Jim and the Indians)

Summary on fuzzies

- "Fuzzies" are our apparently moral feelings that don't follow a utilitarian calculus:
  - Guilt at doing actions that are net-beneficial (Jim and the Indians)
  - Warm feelings from helping people that are disproportionate to benefit (lawyer soup kitchen)

Summary on fuzzies

- "Fuzzies" are our apparently moral feelings that don't follow a utilitarian calculus:
  - Guilt at doing actions that are net-beneficial (Jim and the Indians)
  - Warm feelings from helping people that are disproportionate to benefit (lawyer soup kitchen)
  - Any special concern we feel for those who are close to us, physically or otherwise

Summary on fuzzies

- "Fuzzies" are our apparently moral feelings that don't follow a utilitarian calculus:
  - Guilt at doing actions that are net-beneficial (Jim and the Indians)
  - Warm feelings from helping people that are disproportionate to benefit (lawyer soup kitchen)
  - Any special concern we feel for those who are close to us, physically or otherwise
- Utilitarianism says: we should recognize fuzzies as *mere* feelings, and the moral judgments they express as unreliable.

Summary on fuzzies

- "Fuzzies" are our apparently moral feelings that don't follow a utilitarian calculus:
  - Guilt at doing actions that are net-beneficial (Jim and the Indians)
  - Warm feelings from helping people that are disproportionate to benefit (lawyer soup kitchen)
  - Any special concern we feel for those who are close to us, physically or otherwise
- Utilitarianism says: we should recognize fuzzies as *mere* feelings, and the moral judgments they express as unreliable.
- Accordingly, we should not allow our lives to be directed by our fuzzies.

Summary on fuzzies

- "Fuzzies" are our apparently moral feelings that don't follow a utilitarian calculus:
  - Guilt at doing actions that are net-beneficial (Jim and the Indians)
  - Warm feelings from helping people that are disproportionate to benefit (lawyer soup kitchen)
  - Any special concern we feel for those who are close to us, physically or otherwise
- Utilitarianism says: we should recognize fuzzies as *mere* feelings, and the moral judgments they express as unreliable.
- Accordingly, we should not allow our lives to be directed by our fuzzies.
- We should *manage* our fuzzies by finding easy ways to satisfy them, so that our greater efforts can be reserved for what *really* matters morally.

A partial diagnosis

- Imagine reading Eliezer's posts as someone who isn't getting any fuzzies.

A partial diagnosis

- Imagine reading Eliezer's posts as someone who isn't getting any fuzzies.
- For example:

A partial diagnosis

- Imagine reading Eliezer's posts as someone who isn't getting any fuzzies.
- For example:
  - You're a very lonely person stuck online, or

A partial diagnosis

- Imagine reading Eliezer's posts as someone who isn't getting any fuzzies.
- For example:
  - You're a very lonely person stuck online, or
  - You work in an alienating job that has no obvious connection to doing anything good for anyone

A partial diagnosis

- Imagine reading Eliezer's posts as someone who isn't getting any fuzzies.
- For example:
  - You're a very lonely person stuck online, or
  - You work in an alienating job that has no obvious connection to doing anything good for anyone
- Maybe on some level you miss those fuzzies, resent the people who get them, and resent the public culture that says being a good person is all about the fuzzies.

A partial diagnosis

- Imagine reading Eliezer's posts as someone who isn't getting any fuzzies.
- For example:
  - You're a very lonely person stuck online, or
  - You work in an alienating job that has no obvious connection to doing anything good for anyone
- Maybe on some level you miss those fuzzies, resent the people who get them, and resent the public culture that says being a good person is all about the fuzzies.
- You will be reassured by a moral theory that comes up with a silly little name for fuzzies and says they're bullshit.

A partial diagnosis

- Imagine reading Eliezer's posts as someone who isn't getting any fuzzies.
- For example:
  - You're a very lonely person stuck online, or
  - You work in an alienating job that has no obvious connection to doing anything good for anyone
- Maybe on some level you miss those fuzzies, resent the people who get them, and resent the public culture that says being a good person is all about the fuzzies.
- You will be reassured by a moral theory that comes up with a silly little name for fuzzies and says they're bullshit.
- Of course, none of this means the anti-fuzzy arguments are wrong.

Effectiveness research in EA

- Instead of pursuing fuzzies we should "shut up and multiply," meaning:

Effectiveness research in EA

- Instead of pursuing fuzzies we should "shut up and multiply," meaning:
  - Make serious attempts to calculate the expected benefits of different interventions, and choose the ones with the largest benefits.

Effectiveness research in EA

- Instead of pursuing fuzzies we should "shut up and multiply," meaning:
  - Make serious attempts to calculate the expected benefits of different interventions, and choose the ones with the largest benefits.
- Since 2009 or so, a growing "Effective Altruism" movement has tried to put this prescription into practice.

Effectiveness research in EA

- Instead of pursuing fuzzies we should "shut up and multiply," meaning:
  - Make serious attempts to calculate the expected benefits of different interventions, and choose the ones with the largest benefits.
- Since 2009 or so, a growing "Effective Altruism" movement has tried to put this prescription into practice.
- Effectiveness research is a very good idea and EA efforts have probably saved thousands of lives.

“Earning to Give”

- The fuzziness of your job is less important than how much you will be able to help people in your job.

“Earning to Give”

- The fuzziness of your job is less important than how much you will be able to help people in your job.
- Accordingly, you should maximize your income by working at Goldman Sachs so you can do the most good through your donations.

What if anti-fuzzy signaling is more important than actual effectiveness?

> *[GiveWell's] website juiced donors by advertising its "in-depth evaluations" of "highly effective charities" which do "an incredible amount of good." The pitch came with precise figures . . . .*
>
> *GiveWell's "indepth research" found [a deworming charity] "highly effective." Yet what was GiveWell's "strongest piece of evidence" that the charity improved on what local governments were doing anyway? [A] single interview with a low-level official in one of the five countries where the charity worked. . . .*
>
> *[T]he calculations are hedged with phrases like "very rough guess," "very limited data," "we don't feel confident," "we are highly uncertain," "subjective and uncertain inputs." These pages also say that "we consider our cost-effectiveness numbers to be extremely rough," and that these numbers "should not be taken literally." (Wenar 2024)*

Altruism vs. cooperation

- What if fuzzies were not pure self-indulgence?

Altruism vs. cooperation

- What if fuzzies were not pure self-indulgence?
- EA is implicitly committed to the idea that the greatest good is produced by altruism: some people helping others out of the goodness of their hearts.

Altruism vs. cooperation

- What if fuzzies were not pure self-indulgence?
- EA is implicitly committed to the idea that the greatest good is produced by altruism: some people helping others out of the goodness of their hearts.
- This is distinct from cooperation: people working together without distinguished active and passive participants.

Altruism vs. cooperation

- What if fuzzies were not pure self-indulgence?
- EA is implicitly committed to the idea that the greatest good is produced by altruism: some people helping others out of the goodness of their hearts.
- This is distinct from cooperation: people working together without distinguished active and passive participants.
- If cooperation is actually important, and fuzzies are instrumental to cooperation, then

The hard stuff

The hard stuff

- EA charities have helped people in concrete, non-fantastical ways that are easy to understand, like the bed nets that help prevent malaria.

The hard stuff

- EA charities have helped people in concrete, non-fantastical ways that are easy to understand, like the bed nets that help prevent malaria.
- But "shut up and multiply" means that we shouldn't limit our efforts to the concrete, non-fantastical, and easy to understand.

The hard stuff

- EA charities have helped people in concrete, non-fantastical ways that are easy to understand, like the bed nets that help prevent malaria.
- But "shut up and multiply" means that we shouldn't limit our efforts to the concrete, non-fantastical, and easy to understand.
- Existential risk: a very small risk of everyone dying should be taken just as seriously as a certainty of a small number of people dying.

The hard stuff

- EA charities have helped people in concrete, non-fantastical ways that are easy to understand, like the bed nets that help prevent malaria.
- But "shut up and multiply" means that we shouldn't limit our efforts to the concrete, non-fantastical, and easy to understand.
- Existential risk: a very small risk of everyone dying should be taken just as seriously as a certainty of a small number of people dying.
- Longtermism: just as proximity in space isn't *per se* morally relevant, neither is proximity in time.

The hard stuff

- EA charities have helped people in concrete, non-fantastical ways that are easy to understand, like the bed nets that help prevent malaria.
- But "shut up and multiply" means that we shouldn't limit our efforts to the concrete, non-fantastical, and easy to understand.
- Existential risk: a very small risk of everyone dying should be taken just as seriously as a certainty of a small number of people dying.
- Longtermism: just as proximity in space isn't *per se* morally relevant, neither is proximity in time.
  - If (possibly) there will be a lot of people in the future, and

The hard stuff

- EA charities have helped people in concrete, non-fantastical ways that are easy to understand, like the bed nets that help prevent malaria.
- But "shut up and multiply" means that we shouldn't limit our efforts to the concrete, non-fantastical, and easy to understand.
- Existential risk: a very small risk of everyone dying should be taken just as seriously as a certainty of a small number of people dying.
- Longtermism: just as proximity in space isn't *per se* morally relevant, neither is proximity in time.
  - If (possibly) there will be a lot of people in the future, and
  - we can do things now to make things better for them, then

The hard stuff

- EA charities have helped people in concrete, non-fantastical ways that are easy to understand, like the bed nets that help prevent malaria.
- But "shut up and multiply" means that we shouldn't limit our efforts to the concrete, non-fantastical, and easy to understand.
- Existential risk: a very small risk of everyone dying should be taken just as seriously as a certainty of a small number of people dying.
- Longtermism: just as proximity in space isn't *per se* morally relevant, neither is proximity in time.
  - If (possibly) there will be a lot of people in the future, and
  - we can do things now to make things better for them, then
  - that is just as important as helping people today.

The hard stuff

- EA charities have helped people in concrete, non-fantastical ways that are easy to understand, like the bed nets that help prevent malaria.
- But "shut up and multiply" means that we shouldn't limit our efforts to the concrete, non-fantastical, and easy to understand.
- Existential risk: a very small risk of everyone dying should be taken just as seriously as a certainty of a small number of people dying.
- Longtermism: just as proximity in space isn't *per se* morally relevant, neither is proximity in time.
  - If (possibly) there will be a lot of people in the future, and
  - we can do things now to make things better for them, then
  - that is just as important as helping people today.
  - For a criticism, see Schwitzgebel (2023), "The Washout Argument Against Longtermism."

Existential risk from AI

Existential risk from AI

- One very popular existential risk in the EA community is AI risk.

Existential risk from AI

- One very popular existential risk in the EA community is AI risk.
- Many think the risk of bad outcomes from AI is greater than e.g. that from climate change or nuclear war.

Existential risk from AI

- One very popular existential risk in the EA community is AI risk.
- Many think the risk of bad outcomes from AI is greater than e.g. that from climate change or nuclear war.
- Sketch argument for AI risk (see Bostrom 2012, *Superintelligence*; generally the SF of Vernor Vinge, especially *A Fire Upon the Deep*, 1992; Greg Egan 1995, *Permutation City*):

Existential risk from AI

- One very popular existential risk in the EA community is AI risk.
- Many think the risk of bad outcomes from AI is greater than e.g. that from climate change or nuclear war.
- Sketch argument for AI risk (see Bostrom 2012, *Superintelligence*; generally the SF of Vernor Vinge, especially *A Fire Upon the Deep*, 1992; Greg Egan 1995, *Permutation City*):
  - The pursuit of profit and dominance will lead human actors to develop "artificial general intelligence" (AGI), i.e., AI that can perform most tasks at a human level

Existential risk from AI

- One very popular existential risk in the EA community is AI risk.
- Many think the risk of bad outcomes from AI is greater than e.g. that from climate change or nuclear war.
- Sketch argument for AI risk (see Bostrom 2012, *Superintelligence*; generally the SF of Vernor Vinge, especially *A Fire Upon the Deep*, 1992; Greg Egan 1995, *Permutation City*):
  - The pursuit of profit and dominance will lead human actors to develop "artificial general intelligence" (AGI), i.e., AI that can perform most tasks at a human level
  - AGI will be able to increase its capabilities beyond the human level very quickly (basically computers can expand their capacities faster than humans can)

Existential risk from AI

- One very popular existential risk in the EA community is AI risk.
- Many think the risk of bad outcomes from AI is greater than e.g. that from climate change or nuclear war.
- Sketch argument for AI risk (see Bostrom 2012, *Superintelligence*; generally the SF of Vernor Vinge, especially *A Fire Upon the Deep*, 1992; Greg Egan 1995, *Permutation City*):
  - The pursuit of profit and dominance will lead human actors to develop "artificial general intelligence" (AGI), i.e., AI that can perform most tasks at a human level
  - AGI will be able to increase its capabilities beyond the human level very quickly (basically computers can expand their capacities faster than humans can)
  - The priorities of any AGI will not automatically align with recognizably human values

Existential risk from AI

- One very popular existential risk in the EA community is AI risk.
- Many think the risk of bad outcomes from AI is greater than e.g. that from climate change or nuclear war.
- Sketch argument for AI risk (see Bostrom 2012, *Superintelligence*; generally the SF of Vernor Vinge, especially *A Fire Upon the Deep*, 1992; Greg Egan 1995, *Permutation City*):
  - The pursuit of profit and dominance will lead human actors to develop "artificial general intelligence" (AGI), i.e., AI that can perform most tasks at a human level
  - AGI will be able to increase its capabilities beyond the human level very quickly (basically computers can expand their capacities faster than humans can)
  - The priorities of any AGI will not automatically align with recognizably human values
  - Accordingly, there is a decent chance of superhuman AI that will lead to outcomes like human extinction

The AI researchers' views

- You might think existential risk from AI is sci-fi fluff that real researchers don't take seriously.

The AI researchers' views

- You might think existential risk from AI is sci-fi fluff that real researchers don't take seriously.
- We have surveys on how AI researchers view the potential risks from AI.

The AI researchers' views

- You might think existential risk from AI is sci-fi fluff that real researchers don't take seriously.
- We have surveys on how AI researchers view the potential risks from AI.
  - AI researchers in this context means: authors who publish in NeurIPS and ICML and in some cases other similar conferences

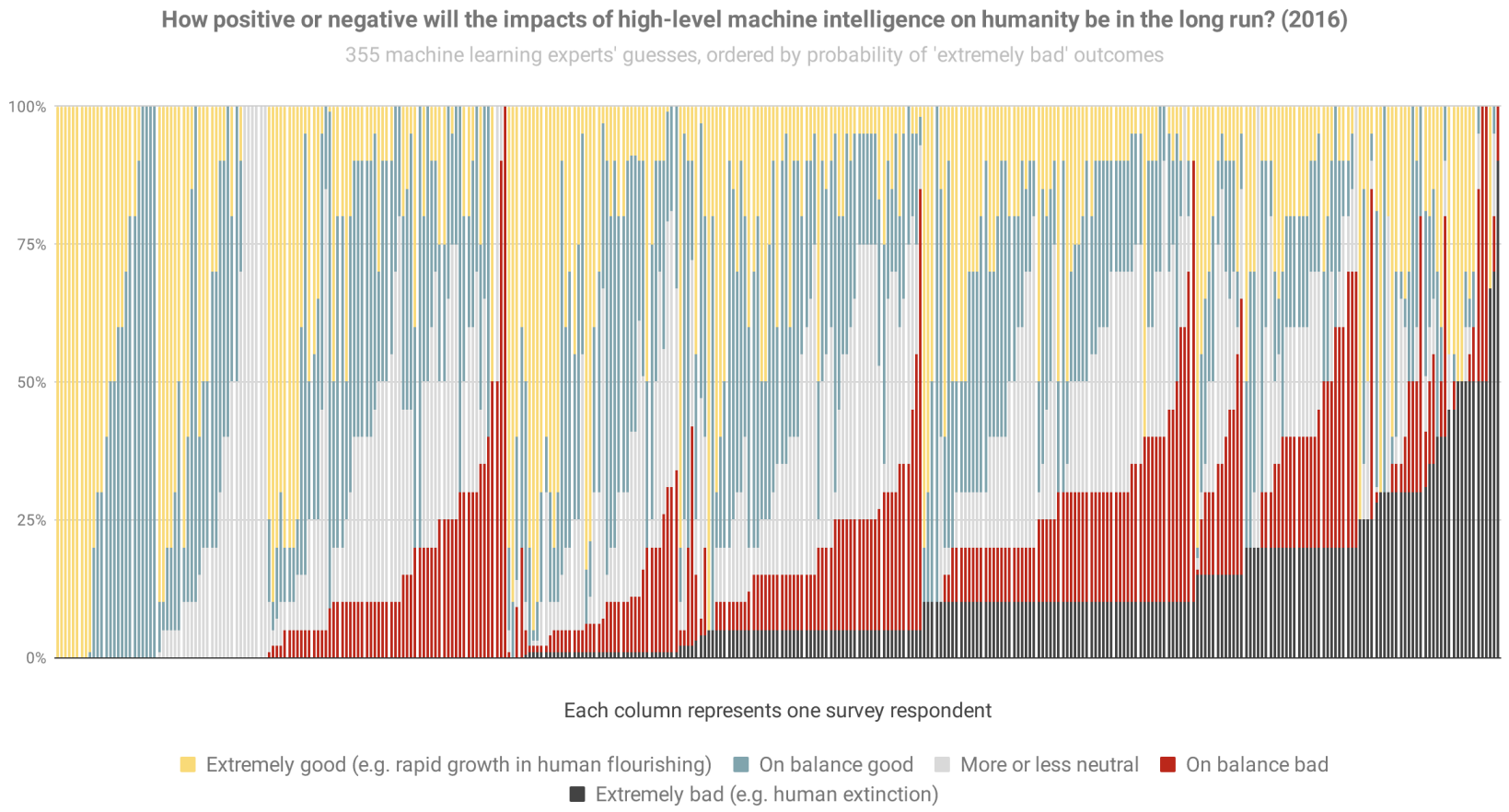
The AI researchers' views

- You might think existential risk from AI is sci-fi fluff that real researchers don't take seriously.
- We have surveys on how AI researchers view the potential risks from AI.
  - AI researchers in this context means: authors who publish in NeurIPS and ICML and in some cases other similar conferences
  - A large share of papers in these conferences come from a small number of companies

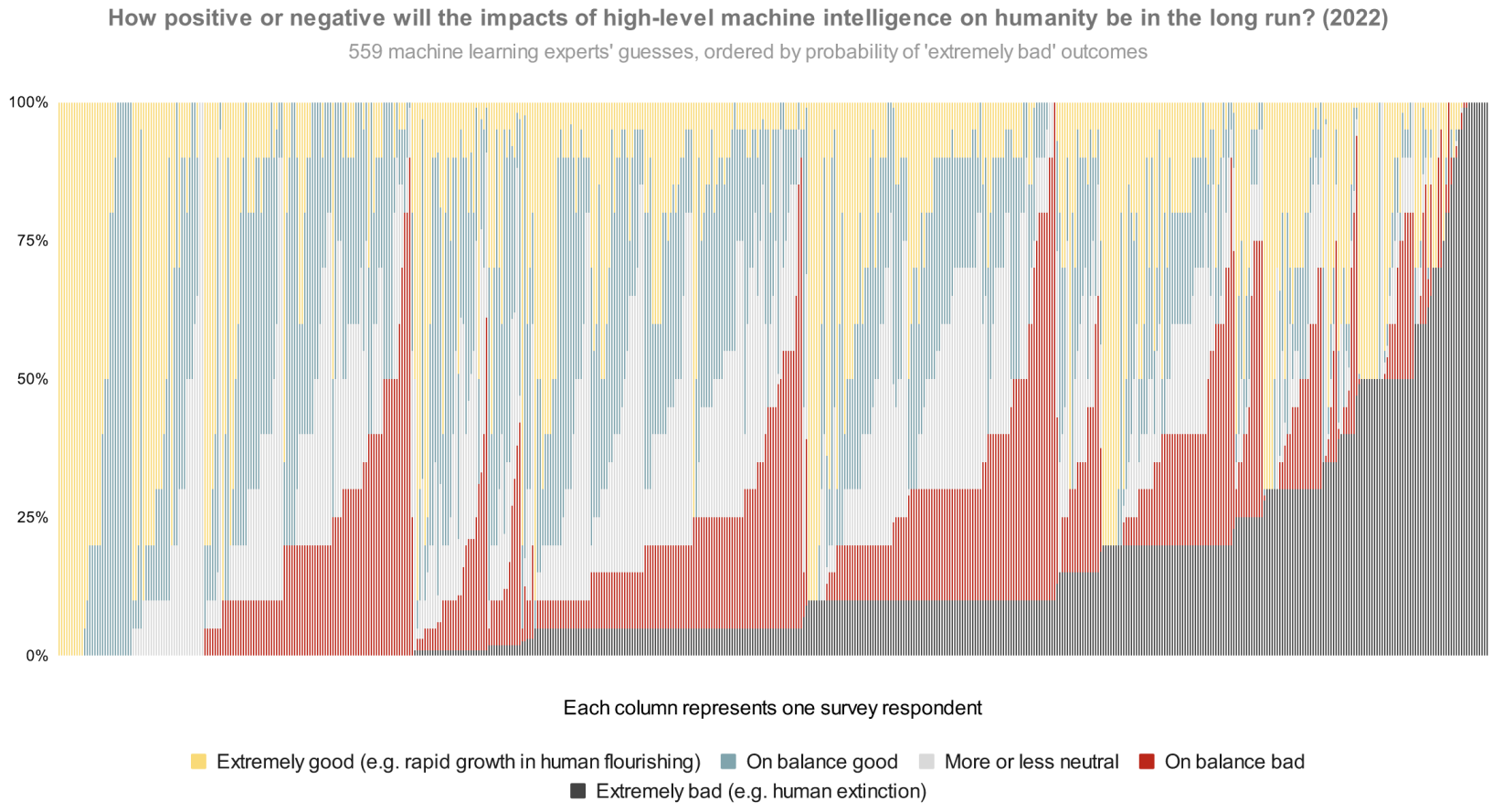
The AI researchers' views

- You might think existential risk from AI is sci-fi fluff that real researchers don't take seriously.
- We have surveys on how AI researchers view the potential risks from AI.
  - AI researchers in this context means: authors who publish in NeurIPS and ICML and in some cases other similar conferences
  - A large share of papers in these conferences come from a small number of companies
  - So the views in the surveys are somewhat representative of the views of researchers in industry

2016 Survey


How positive or negative will the impacts of high-level machine intelligence on humanity be in the long run? (2016)
355 machine learning experts' guesses, ordered by probability of 'extremely bad' outcomes
100%
75%
50%
25%
0%
Each column represents one survey respondent
Extremely good (e.g. rapid growth in human flourishing)
On balance good
More or less neutral
On balance bad
Extremely bad (e.g. human extinction)

2022 Survey


How positive or negative will the impacts of high-level machine intelligence on humanity be in the long run? (2022)
559 machine learning experts' guesses, ordered by probability of 'extremely bad' outcomes
100%
75%
50%
25%
0%
Each column represents one survey respondent
Extremely good (e.g. rapid growth in human flourishing)
On balance good
More or less neutral
On balance bad
Extremely bad (e.g. human extinction)

AI researchers' responses to existential risk

- How it started:

AI researchers' responses to existential risk

- How it started:
  - DeepMind: acquisition by Google conditioned on creating a supervisory AI ethics board

AI researchers' responses to existential risk

- How it started:
  - DeepMind: acquisition by Google conditioned on creating a supervisory AI ethics board
  - OpenAI: organized as a non-profit, with its non-profit board formally in control of its profit-making subsidiary

AI researchers' responses to existential risk

- How it started:
  - DeepMind: acquisition by Google conditioned on creating a supervisory AI ethics board
  - OpenAI: organized as a non-profit, with its non-profit board formally in control of its profit-making subsidiary
  - Anthropic: founded with a commitment not to advance state-of-the-art capabilities

AI researchers' responses to existential risk

- How it started:
  - DeepMind: acquisition by Google conditioned on creating a supervisory AI ethics board
  - OpenAI: organized as a non-profit, with its non-profit board formally in control of its profit-making subsidiary
  - Anthropic: founded with a commitment not to advance state-of-the-art capabilities
- How it's going:

AI researchers' responses to existential risk

- How it started:
  - DeepMind: acquisition by Google conditioned on creating a supervisory AI ethics board
  - OpenAI: organized as a non-profit, with its non-profit board formally in control of its profit-making subsidiary
  - Anthropic: founded with a commitment not to advance state-of-the-art capabilities
- How it's going:
  - DeepMind: nothing is known about the AI ethics board and what it does

AI researchers' responses to existential risk

- How it started:
  - DeepMind: acquisition by Google conditioned on creating a supervisory AI ethics board
  - OpenAI: organized as a non-profit, with its non-profit board formally in control of its profit-making subsidiary
  - Anthropic: founded with a commitment not to advance state-of-the-art capabilities
- How it's going:
  - DeepMind: nothing is known about the AI ethics board and what it does
  - OpenAI: we'll come back to this

AI researchers' responses to existential risk

- How it started:
  - DeepMind: acquisition by Google conditioned on creating a supervisory AI ethics board
  - OpenAI: organized as a non-profit, with its non-profit board formally in control of its profit-making subsidiary
  - Anthropic: founded with a commitment not to advance state-of-the-art capabilities
- How it's going:
  - DeepMind: nothing is known about the AI ethics board and what it does
  - OpenAI: we'll come back to this
  - Anthropic:

AI researchers' responses to existential risk

- How it started:
  - DeepMind: acquisition by Google conditioned on creating a supervisory AI ethics board
  - OpenAI: organized as a non-profit, with its non-profit board formally in control of its profit-making subsidiary
  - Anthropic: founded with a commitment not to advance state-of-the-art capabilities
- How it's going:
  - DeepMind: nothing is known about the AI ethics board and what it does
  - OpenAI: we'll come back to this
  - Anthropic:

Today [March 4, 2024], we're announcing the Claude 3 model family, which sets new industry benchmarks across a wide range of cognitive tasks. The family includes three state-of-the-art models in ascending order of capability: Claude 3 Haiku, Claude 3 Sonnet, and Claude 3 Opus.

The OpenAI putsch

- In November of last year, the OpenAI non-profit board fired its CEO Sam Altman.

The OpenAI putsch

- In November of last year, the OpenAI non-profit board fired its CEO Sam Altman.
  - The reason has never been explained. There is speculation that board members felt that Altman's orientation to the profit-making side was jeopardizing OpenAI's safety mission.

The OpenAI putsch

- In November of last year, the OpenAI non-profit board fired its CEO Sam Altman.
  - The reason has never been explained. There is speculation that board members felt that Altman's orientation to the profit-making side was jeopardizing OpenAI's safety mission.
  - Note that the board had the power and probably the duty to fire the CEO in such a case.

The OpenAI putsch

- In November of last year, the OpenAI non-profit board fired its CEO Sam Altman.
  - The reason has never been explained. There is speculation that board members felt that Altman's orientation to the profit-making side was jeopardizing OpenAI's safety mission.
  - Note that the board had the power and probably the duty to fire the CEO in such a case.
- Then what happened?

The OpenAI putsch

- In November of last year, the OpenAI non-profit board fired its CEO Sam Altman.
  - The reason has never been explained. There is speculation that board members felt that Altman's orientation to the profit-making side was jeopardizing OpenAI's safety mission.
  - Note that the board had the power and probably the duty to fire the CEO in such a case.
- Then what happened?
  - Sam Altman gets on the phone with Microsoft.

The OpenAI putsch

- In November of last year, the OpenAI non-profit board fired its CEO Sam Altman.
  - The reason has never been explained. There is speculation that board members felt that Altman's orientation to the profit-making side was jeopardizing OpenAI's safety mission.
  - Note that the board had the power and probably the duty to fire the CEO in such a case.
- Then what happened?
  - Sam Altman gets on the phone with Microsoft.
  - More than 700 of about 770 employees protest his firing and threaten to move to Microsoft if he is not reinstated.

The OpenAI putsch

- In November of last year, the OpenAI non-profit board fired its CEO Sam Altman.
  - The reason has never been explained. There is speculation that board members felt that Altman's orientation to the profit-making side was jeopardizing OpenAI's safety mission.
  - Note that the board had the power and probably the duty to fire the CEO in such a case.
- Then what happened?
  - Sam Altman gets on the phone with Microsoft.
  - More than 700 of about 770 employees protest his firing and threaten to move to Microsoft if he is not reinstated.
  - Sam Altman comes back and the board is largely replaced

The OpenAI putsch

- In November of last year, the OpenAI non-profit board fired its CEO Sam Altman.
  - The reason has never been explained. There is speculation that board members felt that Altman's orientation to the profit-making side was jeopardizing OpenAI's safety mission.
  - Note that the board had the power and probably the duty to fire the CEO in such a case.
- Then what happened?
  - Sam Altman gets on the phone with Microsoft.
  - More than 700 of about 770 employees protest his firing and threaten to move to Microsoft if he is not reinstated.
  - Sam Altman comes back and the board is largely replaced
- Why didn't this work?

The OpenAI putsch

- In November of last year, the OpenAI non-profit board fired its CEO Sam Altman.
  - The reason has never been explained. There is speculation that board members felt that Altman's orientation to the profit-making side was jeopardizing OpenAI's safety mission.
  - Note that the board had the power and probably the duty to fire the CEO in such a case.
- Then what happened?
  - Sam Altman gets on the phone with Microsoft.
  - More than 700 of about 770 employees protest his firing and threaten to move to Microsoft if he is not reinstated.
  - Sam Altman comes back and the board is largely replaced
- Why didn't this work?
- What could have been done differently?

Kinds of lives

Kinds of lives

- Wiki: “Avraham Marek Klingberg (7 October 1918 – 30 November 2015), known as Marcus Klingberg, was a Polish-born, Israeli epidemiologist and...”

Kinds of lives

- Wiki: “Avraham Marek Klingberg (7 October 1918 – 30 November 2015), known as Marcus Klingberg, was a Polish-born, Israeli epidemiologist and. . . ”
- “[T]he highest ranking Soviet spy ever uncovered in Israel.”

Kinds of lives

- Wiki: “Avraham Marek Klingberg (7 October 1918 – 30 November 2015), known as Marcus Klingberg, was a Polish-born, Israeli epidemiologist and. . . ”
- “[T]he highest ranking Soviet spy ever uncovered in Israel.”
- EA prescriptions call on us to act in ways that seem contradictory:

Kinds of lives

- Wiki: “Avraham Marek Klingberg (7 October 1918 – 30 November 2015), known as Marcus Klingberg, was a Polish-born, Israeli epidemiologist and. . . ”
- “[T]he highest ranking Soviet spy ever uncovered in Israel.”
- EA prescriptions call on us to act in ways that seem contradictory:
  - Do the greatest good by ignoring or minimizing our strongest moral feelings (fuzzies)

Kinds of lives

- Wiki: "Avraham Marek Klingberg (7 October 1918 – 30 November 2015), known as Marcus Klingberg, was a Polish-born, Israeli epidemiologist and..."
- "[T]he highest ranking Soviet spy ever uncovered in Israel."
- EA prescriptions call on us to act in ways that seem contradictory:
  - Do the greatest good by ignoring or minimizing our strongest moral feelings (fuzzies)
  - Help others by working in jobs that don't directly help anyone (earning to give)

Kinds of lives

- Wiki: "Avraham Marek Klingberg (7 October 1918 – 30 November 2015), known as Marcus Klingberg, was a Polish-born, Israeli epidemiologist and..."
- "[T]he highest ranking Soviet spy ever uncovered in Israel."
- EA prescriptions call on us to act in ways that seem contradictory:
  - Do the greatest good by ignoring or minimizing our strongest moral feelings (fuzzies)
  - Help others by working in jobs that don't directly help anyone (earning to give)
  - Work in a field that we believe is being driven to possibly destroy humanity (AI research)

Kinds of lives

- Wiki: "Avraham Marek Klingberg (7 October 1918 – 30 November 2015), known as Marcus Klingberg, was a Polish-born, Israeli epidemiologist and..."
- "[T]he highest ranking Soviet spy ever uncovered in Israel."
- EA prescriptions call on us to act in ways that seem contradictory:
  - Do the greatest good by ignoring or minimizing our strongest moral feelings (fuzzies)
  - Help others by working in jobs that don't directly help anyone (earning to give)
  - Work in a field that we believe is being driven to possibly destroy humanity (AI research)
- To do this is to be something like a spy: to hold yourself apart from your social context and not let it shape your values.