Why We're Doing It and What MIRI Is Doing: Overview and Q&A

Eliezer Yudkowsky

Machine Intelligence Research Institute
intelligence.org

Aug 2, 2014

- Astronomical stakes: $\gg 10^{58}$ QALYs.

- Astronomical stakes: $\gg 10^{58}$ QALYs.
- Intelligence explosion: Sufficiently powerful machine intelligence will improve very rapidly.

- Astronomical stakes: $\gg 10^{58}$ QALYs.
- Intelligence explosion: Sufficiently powerful machine intelligence will improve very rapidly.
- 'Friendly AI'

- Astronomical stakes: $\gg 10^{58}$ QALYs.
- Intelligence explosion: Sufficiently powerful machine intelligence will improve very rapidly.
- 'Friendly AI' (self-improving AI with stable, specifiable, complex, indirect values)

- Astronomical stakes: $\gg 10^{58}$ QALYs.
- Intelligence explosion: Sufficiently powerful machine intelligence will improve very rapidly.
- 'Friendly AI' (self-improving AI with stable, specifiable, complex, indirect values) is
  - Necessary, on pain of losing most QALYs.

- Astronomical stakes: $\gg 10^{58}$ QALYs.
- Intelligence explosion: Sufficiently powerful machine intelligence will improve very rapidly.
- 'Friendly AI' (self-improving AI with stable, specifiable, complex, indirect values) is
  - Necessary, on pain of losing most QALYs.
  - Difficult, not solved by default on current path.

- Astronomical stakes: $\gg 10^{58}$ QALYs.
- Intelligence explosion: Sufficiently powerful machine intelligence will improve very rapidly.
- 'Friendly AI' (self-improving AI with stable, specifiable, complex, indirect values) is
  - Necessary, on pain of losing most QALYs.
  - Difficult, not solved by default on current path.
- Plan: Do direct research, get field on a technical footing, show others technical problems exist.

- Astronomical stakes: $\gg 10^{58}$ QALYs.
- Intelligence explosion: Sufficiently powerful machine intelligence will improve very rapidly.
- 'Friendly AI' (self-improving AI with stable, specifiable, complex, indirect values) is
  - Necessary, on pain of losing most QALYs.
  - Difficult, not solved by default on current path.
- Plan: Do direct research, get field on a technical footing, show others technical problems exist.
- Meta: Plan first on mainline, taking things at face value.

- Astronomical stakes: $\gg 10^{58}$ QALYs.
- Intelligence explosion: Sufficiently powerful machine intelligence will improve very rapidly.
- 'Friendly AI' (self-improving AI with stable, specifiable, complex, indirect values) is
  - Necessary, on pain of losing most QALYs.
  - Difficult, not solved by default on current path.
- Plan: Do direct research, get field on a technical footing, show others technical problems exist.
- Meta: Plan first on mainline, taking things at face value. Worrying about non-face-value scenarios is okay, but make sure face-value scenario is handled.

The future, because it's where the utilons are.

- Reachable stars provide $\gg 10^{58}$ QALYs.

The future, because it's where the utilons are.

- Reachable stars provide $\gg 10^{58}$ QALYs.
- Presentism discredited, future as real as present.

The future, because it's where the utilons are.

- Reachable stars provide $\gg 10^{58}$ QALYs.
- Presentism discredited, future as real as present.
- "Just improve Earth" is one strong candidate.

The future, because it's where the utilons are.

- Reachable stars provide $\gg 10^{58}$ QALYs.
- Presentism discredited, future as real as present.
- "Just improve Earth" is one strong candidate.
- Possible "causal bottlenecks" for direct intervention?

The future, because it's where the utilons are.

- Reachable stars provide $\gg 10^{58}$ QALYs.
- Presentism discredited, future as real as present.
- "Just improve Earth" is one strong candidate.
- Possible "causal bottlenecks" for direct intervention?
  - Extinction risks from synthetic biology?

The future, because it's where the utilons are.

- Reachable stars provide $\gg 10^{58}$ QALYs.
- Presentism discredited, future as real as present.
- "Just improve Earth" is one strong candidate.
- Possible "causal bottlenecks" for direct intervention?
  - Extinction risks from synthetic biology?
  - Molecular nanotechnology & warfare?

The future, because it's where the utilons are.

- Reachable stars provide $\gg 10^{58}$ QALYs.
- Presentism discredited, future as real as present.
- "Just improve Earth" is one strong candidate.
- Possible "causal bottlenecks" for direct intervention?
  - Extinction risks from synthetic biology?
  - Molecular nanotechnology & warfare?
  - Germline engineering causes value drift?

The future, because it's where the utilons are.

- Reachable stars provide $\gg 10^{58}$ QALYs.
- Presentism discredited, future as real as present.
- "Just improve Earth" is one strong candidate.
- Possible "causal bottlenecks" for direct intervention?
  - Extinction risks from synthetic biology?
  - Molecular nanotechnology & warfare?
  - Germline engineering causes value drift?
  - Machine intelligence?

This is not how good it gets.

- Intelligence is incredibly powerful

This is not how good it gets.

- Intelligence is incredibly powerful
- Human brain is not limit

This is not how good it gets.

- Intelligence is incredibly powerful
- Human brain is not limit
  - Firing rate: $\approx$ 100Hz
  - Signal speed: 1-100 meters/sec
  - Heat dissipation: 5e-15 / mm

This is not how good it gets.

- Intelligence is incredibly powerful
- Human brain is not limit
  - Firing rate: $\approx$ 100Hz
  - Signal speed: 1-100 meters/sec
  - Heat dissipation: 5e-15 / mm
  - Software: hacked chimpanzee

This is not how good it gets.

- Intelligence is incredibly powerful
- Human brain is not limit
  - Firing rate: $\approx$ 100Hz
  - Signal speed: 1-100 meters/sec
  - Heat dissipation: 5e-15 / mm
  - Software: hacked chimpanzee
- AI advantages

This is not how good it gets.

- Intelligence is incredibly powerful
- Human brain is not limit
  - Firing rate: $\approx$ 100Hz
  - Signal speed: 1-100 meters/sec
  - Heat dissipation: 5e-15 / mm
  - Software: hacked chimpanzee
- AI advantages
  - Vastly faster serial cognition

This is not how good it gets.

- Intelligence is incredibly powerful
- Human brain is not limit
  - Firing rate: $\approx$ 100Hz
  - Signal speed: 1-100 meters/sec
  - Heat dissipation: 5e-15 / mm
  - Software: hacked chimpanzee
- AI advantages
  - Vastly faster serial cognition
  - Expand to new hardware

This is not how good it gets.

- Intelligence is incredibly powerful
- Human brain is not limit
  - Firing rate: $\approx$ 100Hz
  - Signal speed: 1-100 meters/sec
  - Heat dissipation: 5e-15 / mm
  - Software: hacked chimpanzee
- AI advantages
  - Vastly faster serial cognition
  - Expand to new hardware
  - Rewrite own source code

This is not how fast it gets.

- I.J. Good: Smarter-than-human AI better than humans at building improved AIs.

This is not how fast it gets.

- I.J. Good: Smarter-than-human AI better than humans at building improved AIs.
- Yudkowsky / Chalmers: Threshold of criticality is when δ improvement produces $> \delta$ further improvements
  - This may happen below human level.

This is not how fast it gets.

- I.J. Good: Smarter-than-human AI better than humans at building improved AIs.
- Yudkowsky / Chalmers: Threshold of criticality is when δ improvement produces $> \delta$ further improvements
  - This may happen below human level.
- Diminishing returns?

This is not how fast it gets.

- I.J. Good: Smarter-than-human AI better than humans at building improved AIs.
- Yudkowsky / Chalmers: Threshold of criticality is when δ improvement produces $> \delta$ further improvements
  - This may happen below human level.
- Diminishing returns?
  - Anthropological record of hominids shows linear growth in brain size
  - Above + evolutionary biology implies better software created increasing marginal returns to brain size

This is not how fast it gets.

- I.J. Good: Smarter-than-human AI better than humans at building improved AIs.
- Yudkowsky / Chalmers: Threshold of criticality is when δ improvement produces $> \delta$ further improvements
  - This may happen below human level.
- Diminishing returns?
  - Anthropological record of hominids shows linear growth in brain size
  - Above + evolutionary biology implies better software created increasing marginal returns to brain size
  - Ev-bio also implies minimum fitness return per positive cognitive mutation

This is not how fast it gets.

- I.J. Good: Smarter-than-human AI better than humans at building improved AIs.
- Yudkowsky / Chalmers: Threshold of criticality is when δ improvement produces $> \delta$ further improvements
  - This may happen below human level.
- Diminishing returns?
  - Anthropological record of hominids shows linear growth in brain size
  - Above + evolutionary biology implies better software created increasing marginal returns to brain size
  - Ev-bio also implies minimum fitness return per positive cognitive mutation
- More details: "Intelligence Explosion Microeconomics"

Many possible stable motivations

- Gandhi stability argument

Many possible stable motivations

- Gandhi stability argument: Gandhi doesn't want to self-modify to want to kill people.

Many possible stable motivations

- Gandhi stability argument: Gandhi doesn't want to self-modify to want to kill people.
- Orthogonality thesis

Many possible stable motivations

- Gandhi stability argument: Gandhi doesn't want to self-modify to want to kill people.
- Orthogonality thesis: Can hook up arbitrary preferences to optimization power.

Many possible stable motivations

- Gandhi stability argument: Gandhi doesn't want to self-modify to want to kill people.
- Orthogonality thesis: Can hook up arbitrary preferences to optimization power.
- Space of possible mind designs is very large.

Ethical implications and non-implications

Asserted:

Not asserted:

Ethical implications and non-implications

Asserted:

- Moral internalism is false.

(Moral cognitivism is true.)

Not asserted:

Ethical implications and non-implications

Asserted:

- Moral internalism is false. (Moral cognitivism is true.)

Not asserted:

- Mere machines are morally stupid.

Ethical implications and non-implications

Asserted:

- Moral internalism is false. (Moral cognitivism is true.)
- Only some goal systems imply a wondrously weird intergalactic civilization.

Not asserted:

- Mere machines are morally stupid.

Ethical implications and non-implications

Asserted:

- Moral internalism is false. (Moral cognitivism is true.)
- Only some goal systems imply a wondrously weird intergalactic civilization.

Not asserted:

- Mere machines are morally stupid.
- The galaxies should forever belong to human-sized minds made out of protein.

Ethical implications and non-implications

Asserted:

- Moral internalism is false. (Moral cognitivism is true.)
- Only some goal systems imply a wondrously weird intergalactic civilization.
- It's human brains that carry out our self-judgments of being less than perfect or having room to improve.

Not asserted:

- Mere machines are morally stupid.
- The galaxies should forever belong to human-sized minds made out of protein.

Ethical implications and non-implications

Asserted:

- Moral internalism is false. (Moral cognitivism is true.)
- Only some goal systems imply a wondrously weird intergalactic civilization.
- It's human brains that carry out our self-judgments of being less than perfect or having room to improve.

Not asserted:

- Mere machines are morally stupid.
- The galaxies should forever belong to human-sized minds made out of protein.
- There's no grounds on which to critique any morality, so it's okay for us to stamp our current morals into machines forever.

Why we can't trust debugging, aka "fix it until it stops looking broken"

Why we can't trust debugging, aka "fix it until it stops looking broken"

- Sudden context change

Why we can't trust debugging, aka "fix it until it stops looking broken"

- Sudden context change
- Unforeseen edge-instantiation

Why we can't trust debugging, aka "fix it until it stops looking broken"

- Sudden context change
- Unforeseen edge-instantiation
- Convergent incentive to deceive programmers, resist editing, bide time

What strategy could the Ancient Greeks have used to avoid inevitable moral doom?

What strategy could the Ancient Greeks have used to avoid inevitable moral doom?

- Reflective equilibrium (do what I would mean if I knew more, thought faster, had more self-understanding)

What strategy could the Ancient Greeks have used to avoid inevitable moral doom?

- Reflective equilibrium (do what I would mean if I knew more, thought faster, had more self-understanding)
- This can also avert incentives for present conflict

Starting the projects that people 30 years from now will desperately wish had been started 30 years earlier

- At least 10x as many dung-beetle researchers as Friendly AI researchers.

Starting the projects that people 30 years from now will desperately wish had been started 30 years earlier

- At least 10x as many dung-beetle researchers as Friendly AI researchers.
- Current stage of problem is still running into basic puzzles even assuming unbounded computing power.
- Often long lead-times from foundational math to engineering

Actions and plans

Actions and plans

- Directly challenge the actual hard problems in Friendly AI; turn confusion into math.

Actions and plans

- Directly challenge the actual hard problems in Friendly AI; turn confusion into math.
- Write up technical problems to show other researchers that a technical field exists.

Actions and plans

- Directly challenge the actual hard problems in Friendly AI; turn confusion into math.
- Write up technical problems to show other researchers that a technical field exists.
- Run workshops to bring in outside academics, in cooperation with friendly organizations.

Epistemology and meta

Epistemology and meta

- Any planet that has only 4 full-timers working on Friendly AI, in 2014, is on fire and needs to be put out.

Epistemology and meta

- Any planet that has only 4 full-timers working on Friendly AI, in 2014, is on fire and needs to be put out.
- If what looks like your planet's biggest fire is being ignored, move resources toward putting it out; it's silly to ignore apparent flames just because surface appearances might be wrong.

Epistemology and meta

- Any planet that has only 4 full-timers working on Friendly AI, in 2014, is on fire and needs to be put out.
- If what looks like your planet's biggest fire is being ignored, move resources toward putting it out; it's silly to ignore apparent flames just because surface appearances might be wrong.
- No good ethics without engineering.

Epistemology and meta

- Any planet that has only 4 full-timers working on Friendly AI, in 2014, is on fire and needs to be put out.
- If what looks like your planet's biggest fire is being ignored, move resources toward putting it out; it's silly to ignore apparent flames just because surface appearances might be wrong.
- No good ethics without engineering.
- No good meta-level without object-level.

Q&A time!

- Astronomical stakes: $\gg 10^{58}$ QALYs.
- Intelligence explosion: Sufficiently powerful machine intelligence will improve very rapidly.
- 'Friendly AI'

Q&A time!

- Astronomical stakes: $\gg 10^{58}$ QALYs.
- Intelligence explosion: Sufficiently powerful machine intelligence will improve very rapidly.
- 'Friendly AI' (self-improving AI with stable, specifiable, complex, indirect values) is
  - Necessary, on pain of losing most QALYs.
  - Difficult, not solved by default on current path.
- Plan: Do direct research, get field on a technical footing, show others technical problems exist.
- Meta: Plan first on mainline, taking things at face value. Worrying about non-face-value scenarios is okay, but make sure face-value scenario is handled.