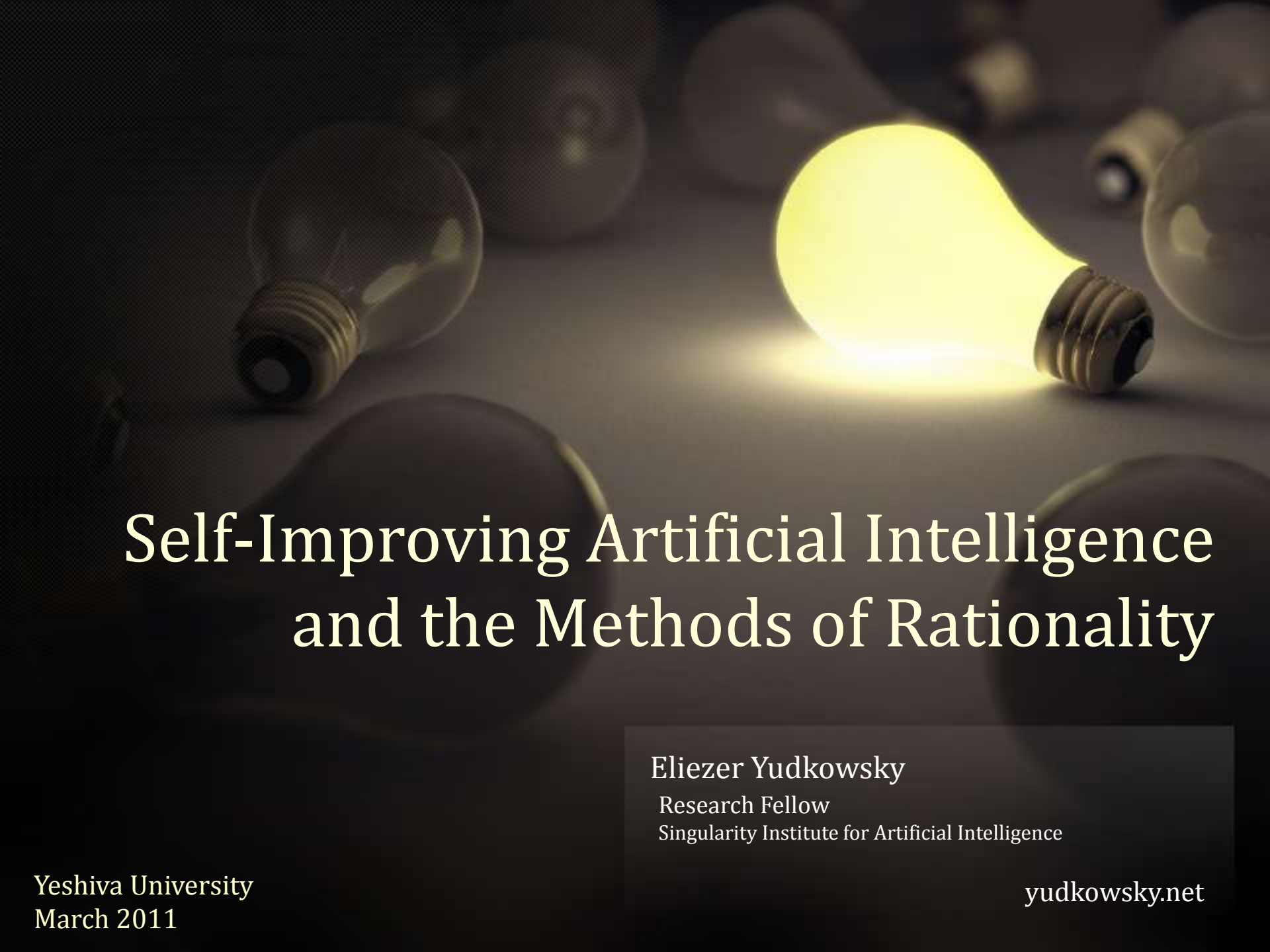


Self-Improving Artificial Intelligence and the Methods of Rationality

Eliezer Yudkowsky
Research Fellow
Singularity Institute for Artificial Intelligence

yudkowsky.net

Yeshiva University
March 2011

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Rank the following statements from most probable to least probable:

- Linda is a teacher in an elementary school
- Linda works in a bookstore and takes Yoga classes
- Linda is active in the feminist movement
- Linda is a psychiatric social worker
- Linda is a member of the League of Women Voters
- Linda is a bank teller
- Linda is an insurance salesperson
- Linda is a bank teller and is active in the feminist movement

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Rank the following statements from most probable to least probable:

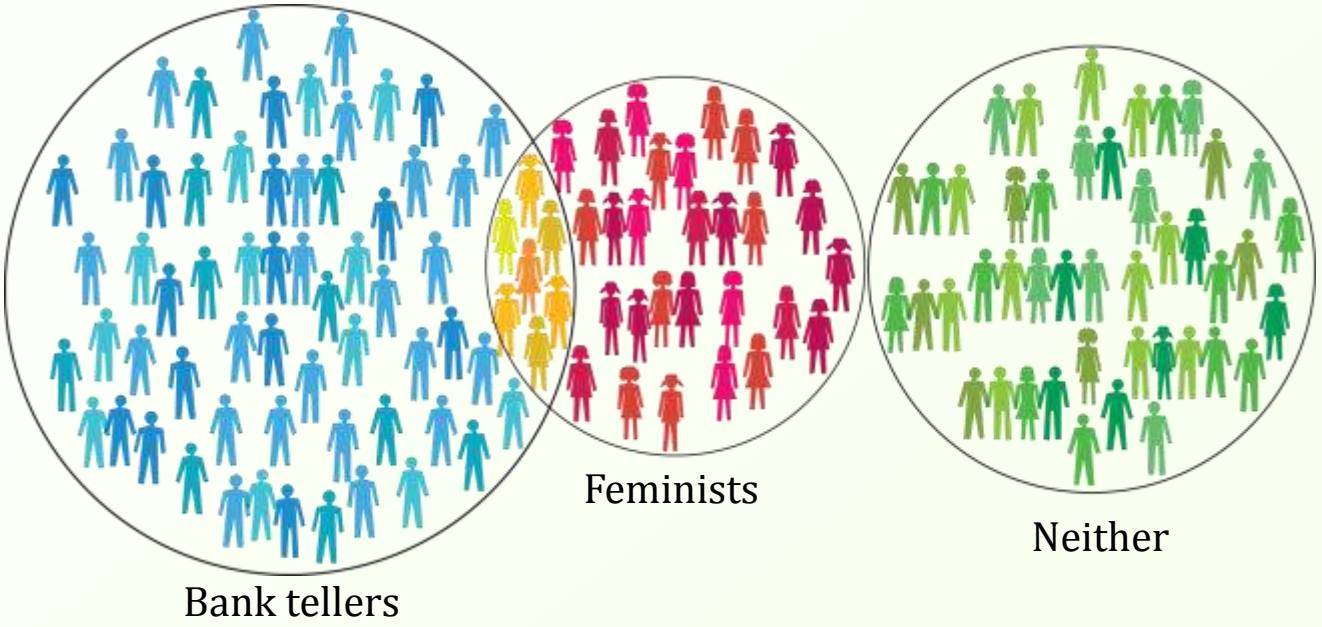
- Linda is a teacher in an elementary school
- Linda works in a bookstore and takes Yoga classes
- Linda is active in the feminist movement
- Linda is a psychiatric social worker
- Linda is a member of the League of Women Voters
- Linda is a bank teller (A)
- Linda is an insurance salesperson
- Linda is a bank teller and is active in the feminist movement (A & B)

89% of subjects thought Linda was more likely to be a feminist bank teller than a bank teller

Conjunction fallacy: $P(A\&B)$ must be $\leq$ than $P(A)$

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

- **Linda is a bank teller (A)**
- **Linda is a bank teller and is active in the feminist movement (A & B)**



WHICH IS THE BEST BET?

- Die with 4 green faces and 2 red faces
- Die rolled 20 times, series of Gs and Rs recorded
- If your chosen sequence appears, you win $25
- Which of these three sequences would you prefer to bet on?

Sequence 1

Sequence 2

Sequence 3

Amos Tversky and Daniel Kahneman (1983), "Extensional versus Intuitive Reasoning". Psychological Review, 90, 293-315

WHICH IS THE BEST BET?

- Die with 4 green faces and 2 red faces
- Die rolled 20 times, series of Gs and Rs recorded
- If your chosen sequence appears, you win $25
- Which of these three sequences would you prefer to bet on?

Sequence 1

Sequence 2

Sequence 3

- 65% chose Sequence 2.
- 125 undergraduates played this gamble with real payoffs.

Amos Tversky and Daniel Kahneman (1983), "Extensional versus Intuitive Reasoning". Psychological Review, 90, 293-315

CONJUNCTION FALLACY: P(A) < P(A&B)

- Die with 4 green faces and 2 red faces
- Die rolled 20 times, series of Gs and Rs recorded
- If your chosen sequence appears, you win $25
- Which of these three sequences would you prefer to bet on?

- but Sequence 1 dominates Sequence 2;
  it appears within 2, so it appears anywhere 2 appears.

Amos Tversky and Daniel Kahneman (1983), "Extensional versus Intuitive Reasoning". Psychological Review, 90, 293-315

CONJUNCTION FALLACY: P(A) < P(A&B)

- Die with 4 green faces and 2 red faces: GGRGGR

Sequence 1

Sequence 2

Sequence 3

Judgment of probability versus Judgment of representativeness

We want: We ask:

Which is most probable? Which is most representative?

Amos Tversky and Daniel Kahneman (1983), "Extensional versus Intuitive Reasoning". Psychological Review, 90, 293-315

2nd INTERNATIONAL CONGRESS ON FORECASTING

- Two independent groups of professional analysts asked:
  Please rate the probability that the following event will occur in 1983.

Version 1

A complete suspension of diplomatic relations between the USA and the Soviet Union, sometime in 1983.

Version 2

A Russian invasion of Poland, and a complete suspension of diplomatic relations between the USA and the Soviet Union, sometime in 1983.

- The group asked to rate Version 2 responded with significantly higher probabilities.

Amos Tversky and Daniel Kahneman (1983), "Extensional versus Intuitive Reasoning". Psychological Review, 90, 293-315

Second International Congress on Forecasting

- Two independent groups of professional analysts were asked to:
- *Please rate the probability that the following event will occur in 1983.*

Version 1

A complete suspension of diplomatic relations between the USA and the Soviet Union, sometime in 1983.

Version 2

A Russian invasion of Poland, and a complete suspension of diplomatic relations between the USA and the Soviet Union, sometime in 1983.

- The group asked to rate Version 2 responded with significantly higher probabilities.

Amos Tversky and Daniel Kahneman (1983), "Extensional versus Intuitive Reasoning". Psychological Review, 90, 293-315

Conjunction Fallacy: P(A) < P(A&B)

- Two independent groups of professional analysts were asked to:
- *Please rate the probability that the following event will occur in 1983.*

Version 1

A complete suspension of diplomatic relations between the USA and the Soviet Union, sometime in 1983.

Version 2

A Russian invasion of Poland, and a complete suspension of diplomatic relations between the USA and the Soviet Union, sometime in 1983.

- The group asked to rate Version 2 responded with significantly higher probabilities.
- Adding detail to a story can make the story sound more plausible, even though the story necessarily becomes less probable.

INSENSITIVITY TO PREDICTABILITY

- Subjects presented with description of teacher during lesson
- Some subjects asked to evaluate quality of lesson, in percentile score
- Other subjects asked to predict percentile standing of teacher 5 years later

Kahneman, D., Tversky, A. 1973. On the psychology of prediction. Psychological Review 80: 237-51.

INSENSITIVITY TO PREDICTABILITY

- Subjects presented with description of teacher during lesson
- Some subjects asked to evaluate quality of lesson, in percentile score
- Other subjects asked to predict percentile standing of teacher 5 years later
- Judgments identical - equally extreme!

Kahneman, D., Tversky, A. 1973. On the psychology of prediction. Psychological Review 80: 237-51.

What do you think you know,
and how do you think you know it?

One of these technologies is not like the others...

Artificial Intelligence	Interplanetary travel
Cancer cure	Nano-manufacturing

The power of intelligence:

- Fire
- Language
- Nuclear weapons
- Skyscrapers
- Spaceships
- Money
- Science

"Book smarts" *vs.* cognition:

"Book smarts" evokes images of:

- Calculus
- Chess
- Good recall of facts

Other stuff that happens *in the brain*:

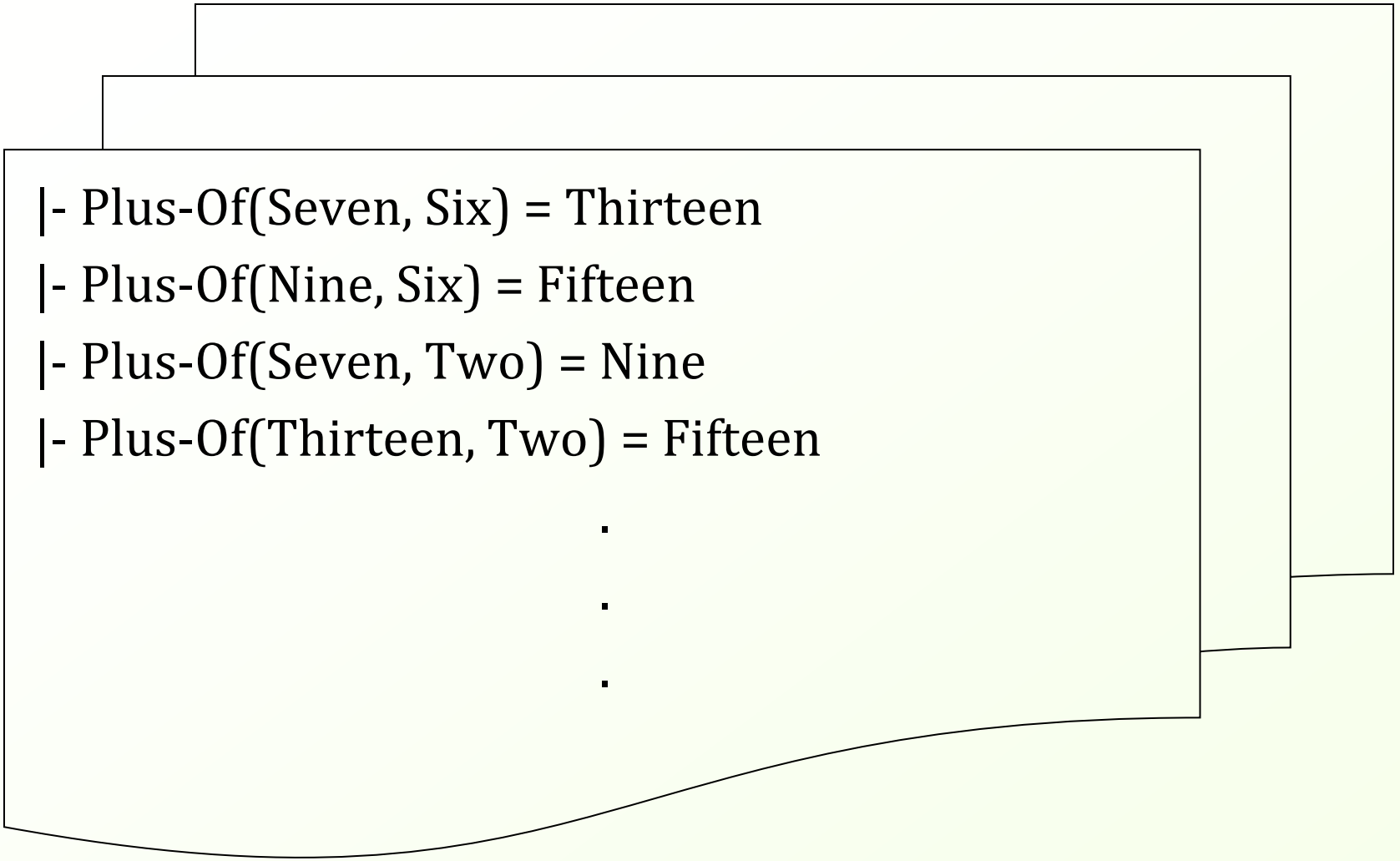
- Social persuasion
- Enthusiasm
- Reading faces
- Rationality
- Strategic cleverness

What do we think we can guess?

- Technologies which *impact upon cognition* will end up mattering most, because intelligence is more powerful and significant than cool devices.

What do you think you know,
and how do you think you know it?

Artificial Addition



Views on Artificial General Addition

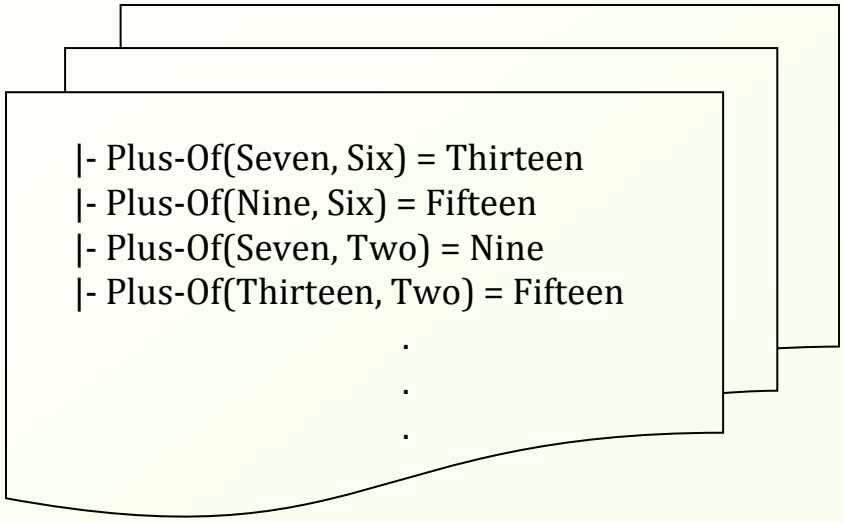
- Framing problem - what 'twenty-one plus' equals depends on whether it's 'plus three' or 'plus four'. Need to program huge network of arithmetical facts to cover common-sense truths.
- Need Artificial Arithmetician that can understand natural language, so instead of being told that twenty-one plus sixteen equals thirty-seven, it can obtain knowledge by reading Web.
- Need to develop General Arithmetician the same way Nature did - evolution.
- Top-down approaches have failed to produce arithmetic. Need bottom-up approach to make arithmetic *emerge.* Accept unpredictability of complex systems.

Views on Artificial General Addition

- Neural networks - just like the human brain! Can be trained without understanding how they work! NNs will do arithmetic without us, their creators, ever understanding how they add.
- Need calculators as powerful as a human brain. Moore's Law predicts availability on April 27, 2031 between 4 and 4:30AM.
- Must simulate neural circuitry humans use for addition.
- Godel's Theorem shows no formal system can ever capture properties of arithmetic. Classical physics is formalizable. Hence AGA must exploit quantum gravity.
- If human arithmetic were simple enough to program, we couldn't count high enough to build computers.
- Haven't you heard of John Searle's Chinese Calculator Experiment?
- Will never know nature of arithmetic; problem is just too hard.

Artificial Addition: The Moral

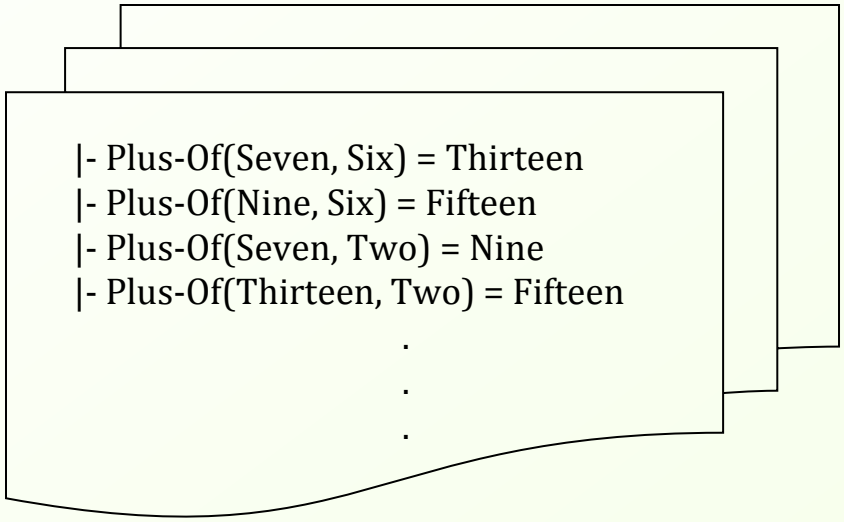
- *Moral 1:* When you're missing a basic insight, you *must* find that insight. Workarounds may *sound* clever, but aren't. Can't talk sensibly about solutions until no-longer-confused.



Artificial Addition: The Moral

- *Moral 1:* When you're missing a basic insight, you *must* find that insight. Workarounds may *sound* clever, but aren't. Can't talk sensibly about solutions until no-longer-confused.

- *Moral 2:* Patching an infinite number of surface cases means you didn't understand the underlying generator.



"The influence of animal or vegetable life on matter is infinitely beyond the range of any scientific inquiry hitherto entered on. Its power of directing the motions of moving particles, in the demonstrated daily miracle of our human free-will, and in the growth of generation after generation of plants from a single seed, are infinitely different from any possible result of the fortuitous concurrence of atoms... Modern biologists were coming once more to the acceptance of something and that was a vital principle."

- Lord Kelvin

Mind Projection Fallacy:

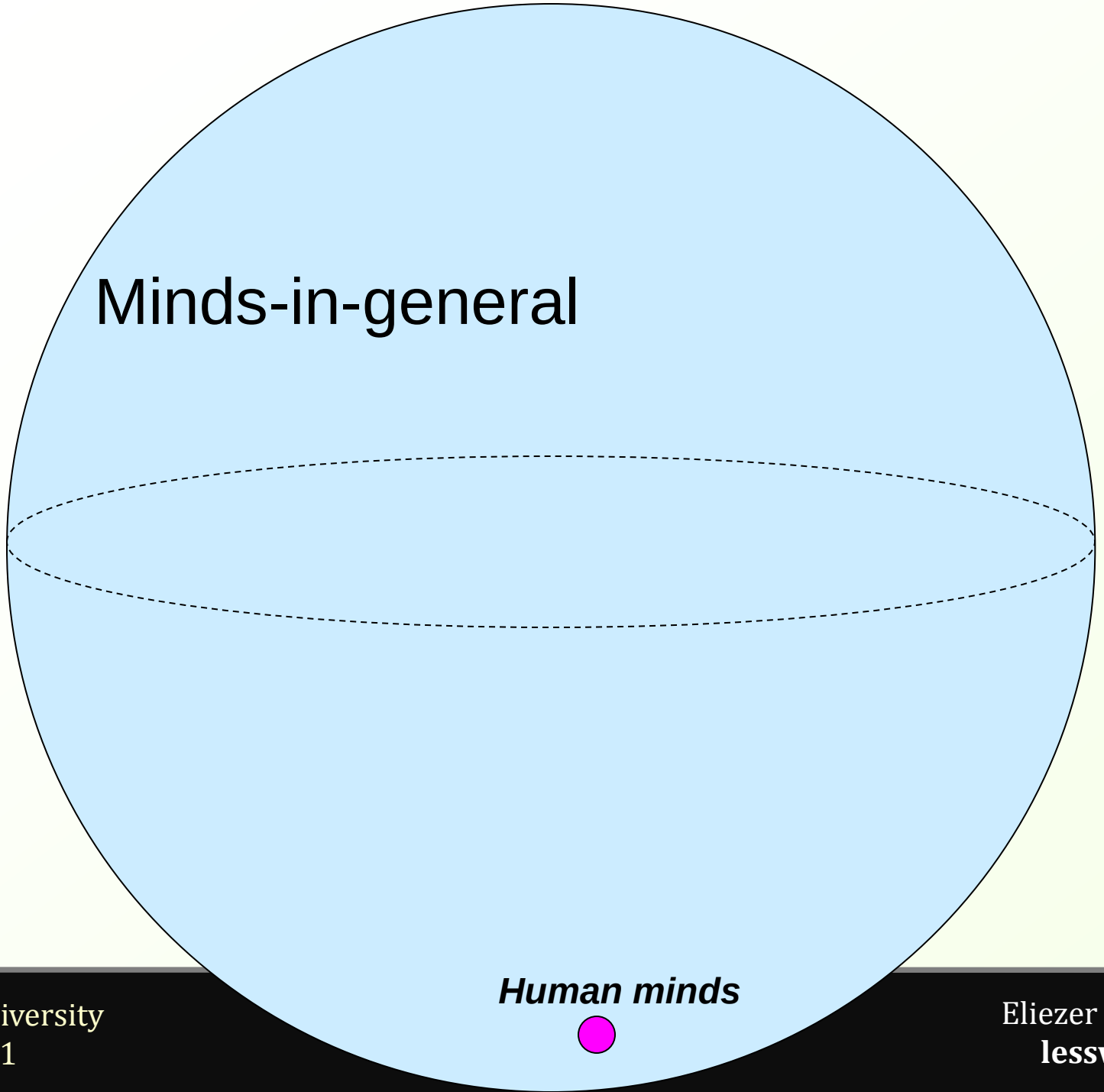
If I am ignorant about a phenomenon,
this is a fact about my state of mind,
not a fact about the phenomenon.

(Jaynes, E.T. 2003. *Probability Theory: The Logic of Science.* Cambridge: Cambridge University Press.)

What do we think we can guess?

- Technologies which *impact upon cognition* will end up mattering most.
- Today's difficulties in constructing AI are not because intelligence is *inherently mysterious,* but because we're currently still ignorant of some basic insights.

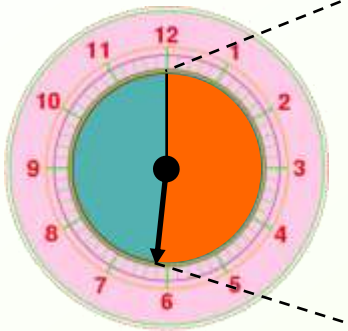
What do you think you know,
and how do you think you know it?


Minds-in-general
Human minds

Biology *not* near top of scale

- Lightspeed >10^6 times faster than axons, dendrites.
- Synaptic spike dissipates >10^6 minimum heat (though transistors do worse)
- Transistor clock speed >>10^6 times faster than neuron spiking frequency

1,000,000-fold speedup physically possible:
(1 year = 31,556,926 seconds) → 31 seconds



2007

January

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

February

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28			

March

				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

April

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

May

		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

June

					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

July

1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

August

			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

September

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

October

	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

November

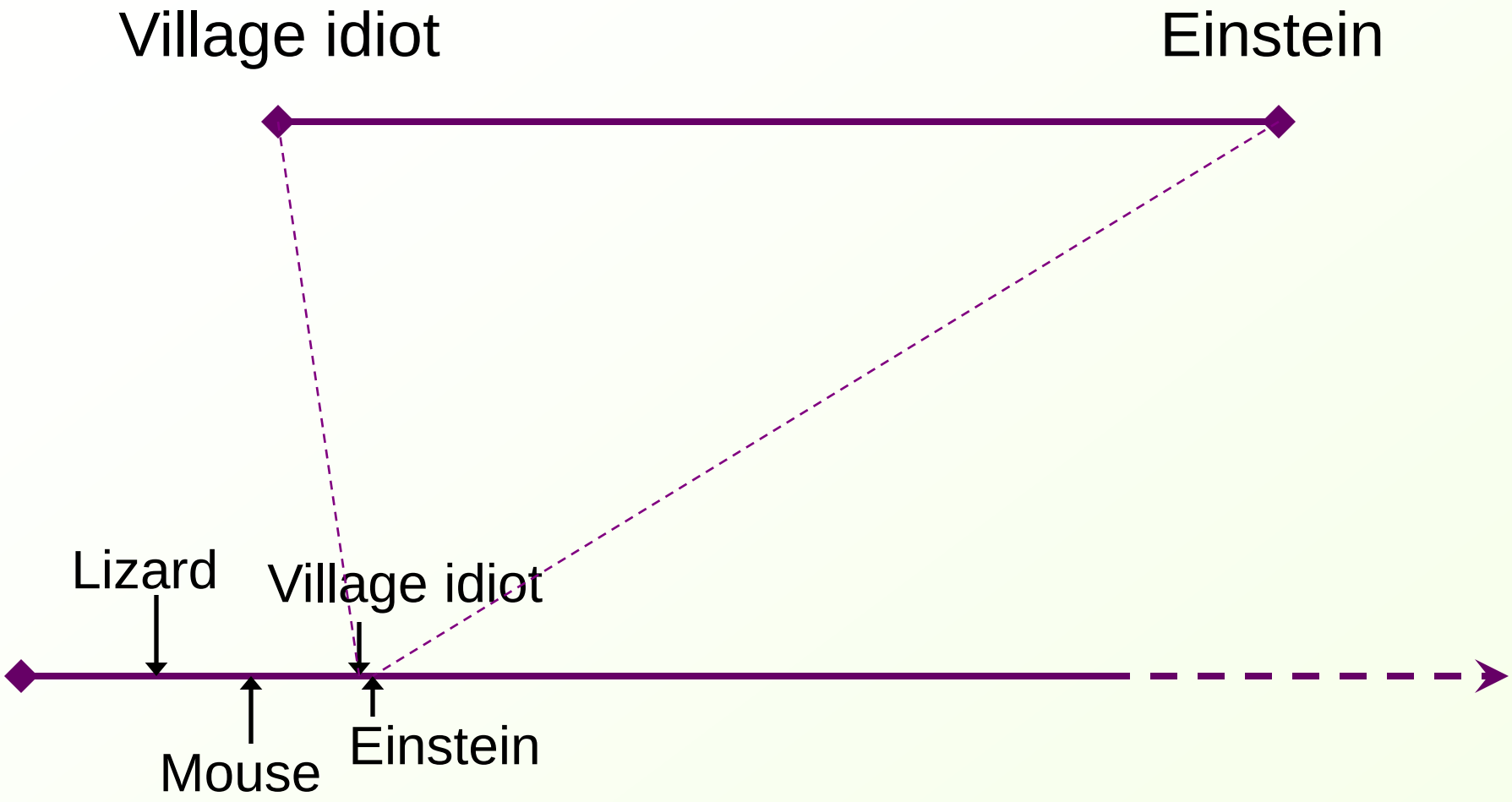
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

December

						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

Village idiot

Einstein


Village idiot
Einstein
Lizard
Village idiot
Mouse
Einstein

AI advantages

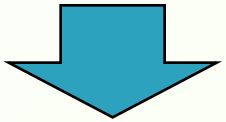
- Total read/write access to own state
- Absorb more hardware
  (possibly orders of magnitude more!)
- Understandable code
- Modular design
- Clean internal environment

What do we think we can guess?

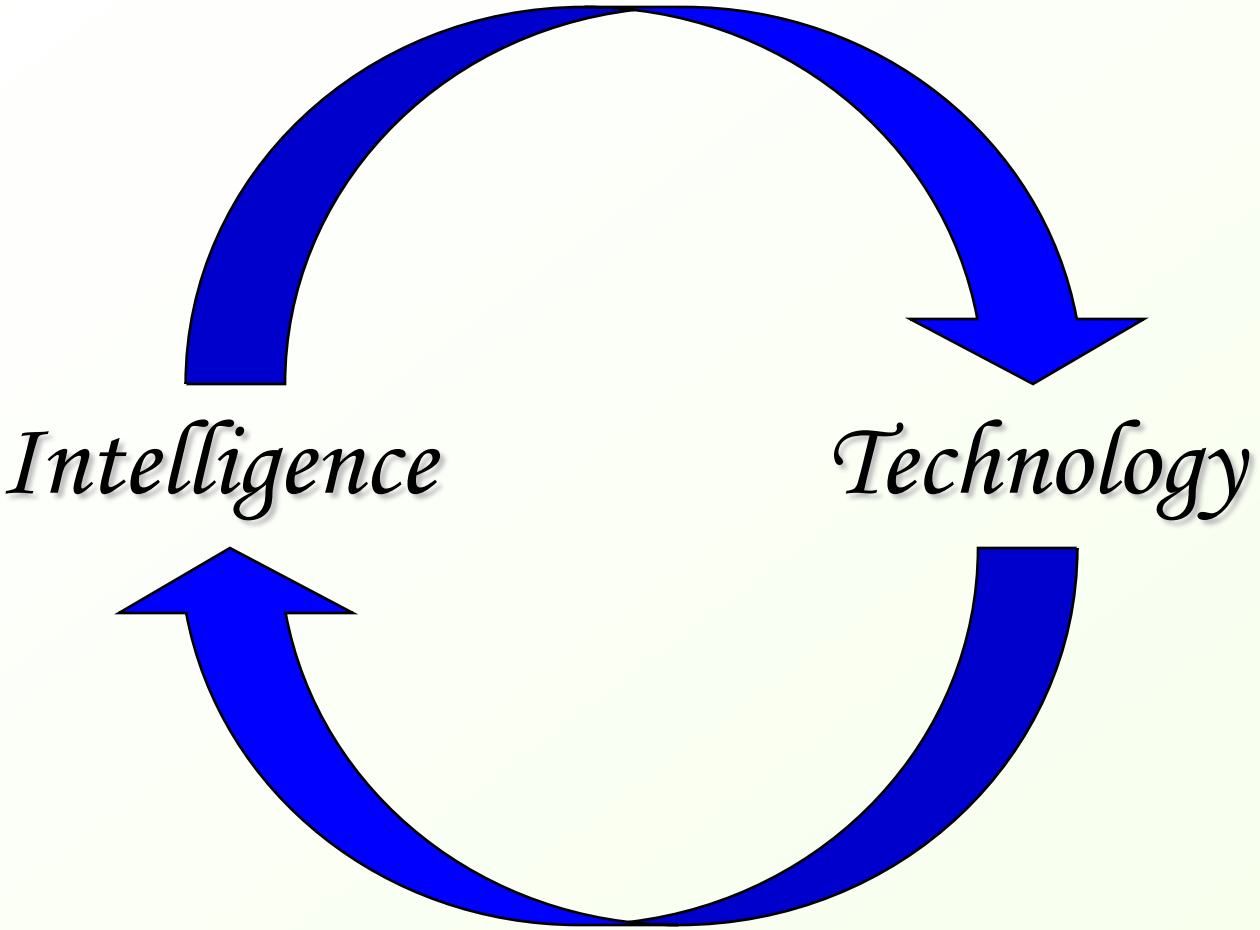
- Technologies which *impact upon cognition* will end up mattering most.
- Today's difficulties in constructing AI are due to ignorance of key insights, not inherent mysteriousness.
- It is possible for minds to exist that are *far* more powerful than human. We can see that human *hardware* is far short of the physical limits, and there's no reason to expect the software to be optimal either.

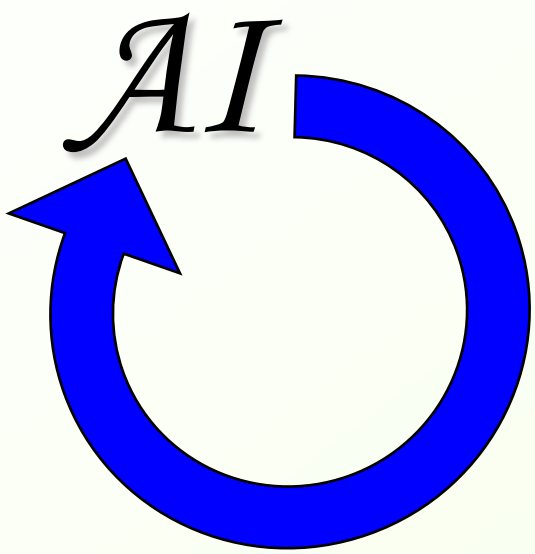
What do you think you know,
and how do you think you know it?

Intelligence



Technology


Intelligence
Technology


AI

"Intelligence explosion:"

- Invented by I. J. Good in 1965
- Hypothesis: More intelligence -> more creativity on task of making yourself even smarter
- Prediction: Positive feedback cycle rapidly creates *superintelligence*.

(Good, I. J. 1965. Speculations Concerning the First Ultraintelligent Machine. Pp. 31-88 in *Advances in Computers*, **6**, F. L. Alt and M. Rubinoff, eds. New York: Academic Press.)

How fast an intelligence explosion?

- From chimpanzees to humans – 4x brain size, 6x frontal cortex, no apparent obstacles to constant evolutionary pressure.
- From flint handaxes to Moon rockets – constant human brains don't slow down over course of technological history as problems get harder.
- Thus an *increasingly* powerful optimization process should undergo *explosive* self-improvement.

Intelligence explosion hypothesis does ***not*** imply, nor require:

- More change occurred from 1970 to 2000 than from 1940 to 1970.

Intelligence explosion hypothesis does ***not*** imply, nor require:

- More change occurred from 1970 to 2000 than from 1940 to 1970.
- Technological progress follows a predictable curve.

What do we think we can guess?

- Technologies which *impact upon cognition* will end up mattering most.
- Today's difficulties in constructing AI are due to ignorance of key insights, not inherent mysteriousness.
- It is possible for minds to exist that are *far* more powerful than human.
- An AI over some threshold level of intelligence will *recursively* improve itself and explode into superintelligence.

What do you think you know,
and how do you think you know it?

In Every Known Human Culture:

- tool making
- weapons
- grammar
- tickling
- meal times
- mediation of conflicts
- dance, singing
- personal names
- promises
- mourning

(Donald E. Brown, 1991. *Human universals.* New York: McGraw-Hill.)

A complex adaptation must be universal within a species.

(John Tooby and Leda Cosmides, 1992. *The Psychological Foundations of Culture.* In *The Adapted Mind,* eds. Barkow, Cosmides, and Tooby.)

A complex adaptation must be universal within a species.

If: 6 necessary genes

Each at 10% frequency in population

Then: 1 in 1,000,000 have complete adaptation

(John Tooby and Leda Cosmides, 1992. *The Psychological Foundations of Culture.* In *The Adapted Mind,* eds. Barkow, Cosmides, and Tooby.)

Incremental evolution of complexity:

1. [A] A is advantageous by itself.
2. [A←B] B depends on A.
3. [A'↔B] A' replaces A, depends on B.
4. [A'B←C] C depends on A' and B

...

(John Tooby and Leda Cosmides, 1992. *The Psychological Foundations of Culture.* In *The Adapted Mind,* eds. Barkow, Cosmides, and Tooby.)

The Psychic Unity of Humankind

Complex adaptations
must be universal in a species –
including *cognitive* machinery
in *Homo sapiens!*

(John Tooby and Leda Cosmides, 1992. *The Psychological Foundations of Culture.* In *The Adapted Mind,* eds. Barkow, Cosmides, and Tooby.)


Must… not…
emote…

The Great Failure of Imagination: *Anthropomorphism*

Yeshiva University
March 2011

Eliezer Yudkowsky
lesswrong.com

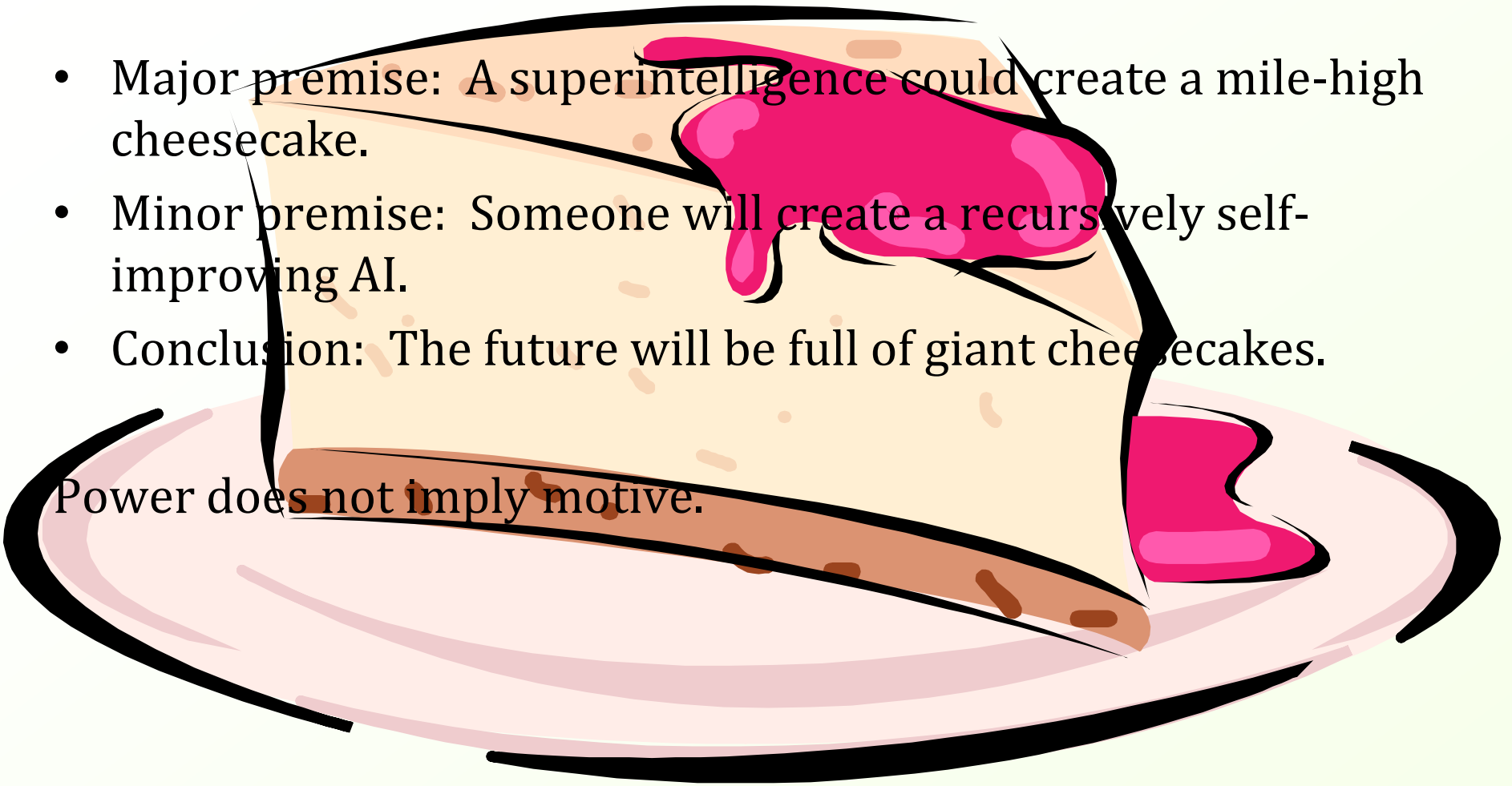

GIANTS AND GIRLS

Fallacy of the Giant Cheesecake

- Major premise: A superintelligence could create a mile-high cheesecake.
- Minor premise: Someone will create a recursively self-improving AI.
- Conclusion: The future will be full of giant cheesecakes.

Power does not imply motive.

Fallacy of the Giant Cheesecake



- Major premise: A superintelligence could create a mile-high cheesecake.
- Minor premise: Someone will create a recursively self-improving AI.
- Conclusion: The future will be full of giant cheesecakes.

Power does not imply motive.

What do we think we can guess?

- Technologies which *impact upon cognition* will end up mattering most.
- Today's difficulties in constructing AI are due to ignorance of key insights, not inherent mysteriousness.
- It is possible for minds to exist that are *far* more powerful than human.
- An AI over some threshold level of intelligence will recursively self-improve and explode into superintelligence.
- Features of "human nature" that we take for granted are just one of a vast number of possibilities, and not all possible agents in that space are friendly.

What do you think you know,
and how do you think you know it?

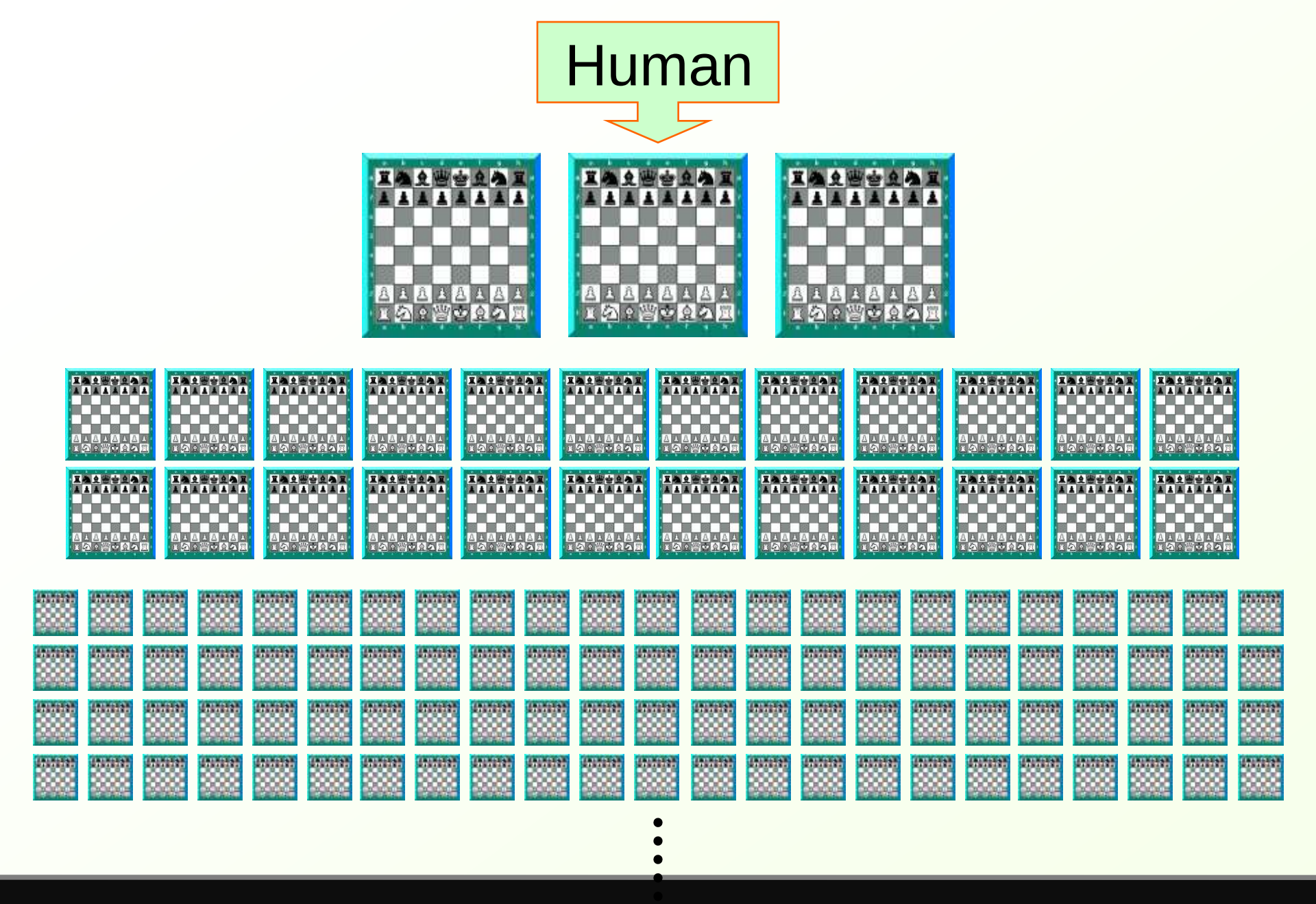
What do you think you know,
and how do you think you know it?

To predict a smarter mind,
you'd have to be that smart yourself?

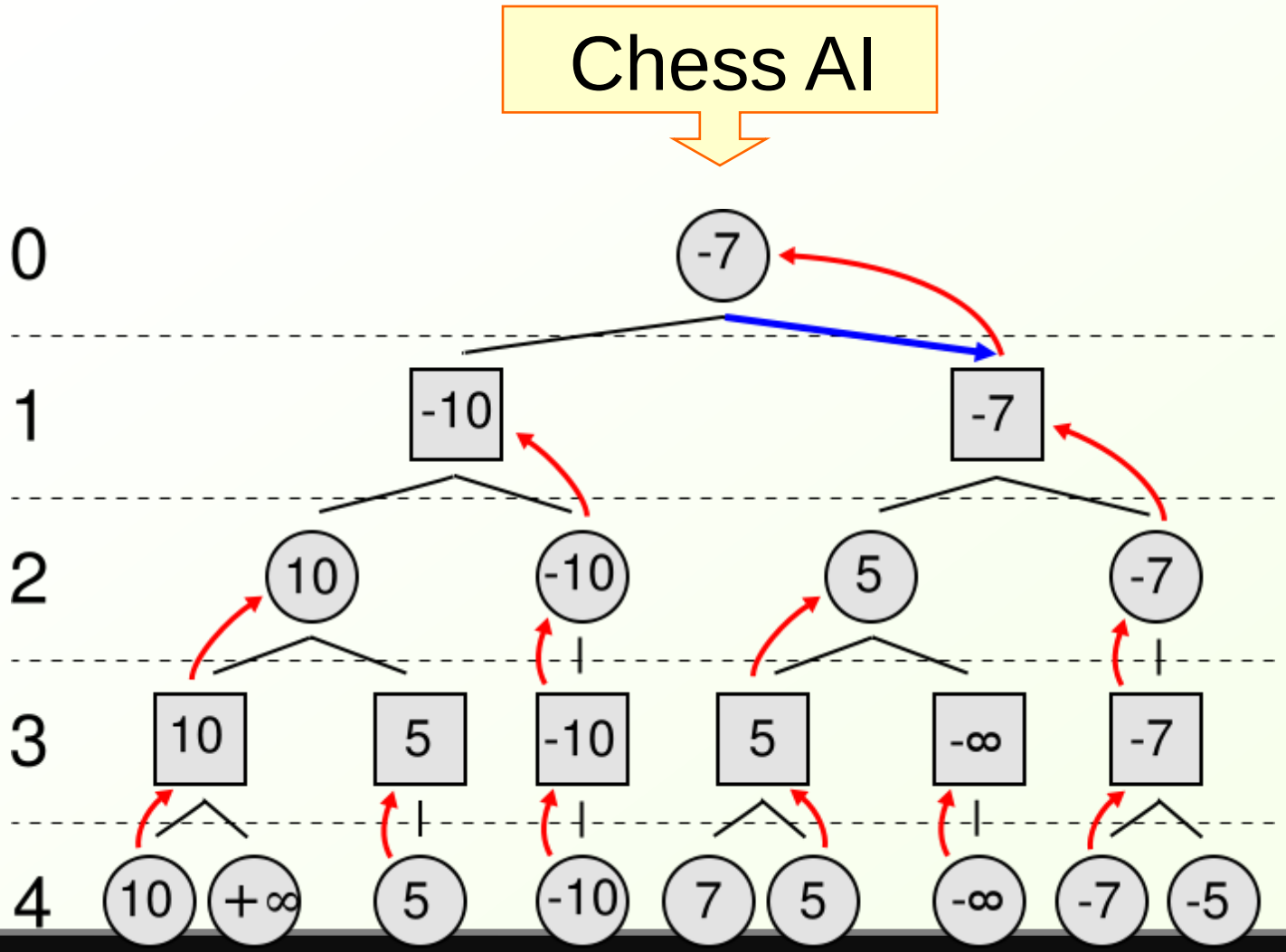
Human

Chess AI

?


Human

Key insight: Predict & value consequences



"To predict a smarter mind, you'd have to be that smart yourself"?

- Deep Blue's programmers couldn't predict its exact moves

"To predict a smarter mind, you'd have to be that smart yourself"?

- Deep Blue's programmers couldn't predict its exact moves
- So why not use random move generator?

"To predict a smarter mind, you'd have to be that smart yourself"?

- Deep Blue's programmers couldn't predict its exact moves
- So why not use random move generator?
- Unpredictability of superior intelligence ≠ unpredictability of coinflips

"To predict a smarter mind, you'd have to be that smart yourself"?

- Deep Blue's programmers couldn't predict its exact moves
- So why not use random move generator?
- Unpredictability of superior intelligence ≠ unpredictability of coinflips
- Deep Blue's "unpredictable" move, predictably has *consequence* of winning game

"To predict a smarter mind, you'd have to be that smart yourself"?

- Deep Blue's programmers couldn't predict its exact moves
- So why not use random move generator?
- Unpredictability of superior intelligence ≠ unpredictability of coinflips
- Deep Blue's "unpredictable" move, predictably has *consequence* of winning game
- Inspection of code can't *prove* consequences in an uncertain real-world environment, but it can establish with near-certainty that an agent is *trying to find the most-probably-good action.*

Stability of goals in self-modifying agents:

- Gandhi doesn't want to kill people.
- Offer Gandhi a pill that will make him enjoy killing people?

Stability of goals in self-modifying agents:

- Gandhi doesn't want to kill people.
- Offer Gandhi a pill that will make him enjoy killing people?

- If Gandhi correctly assesses this is what the pill does, Gandhi will refuse the pill, because the current Gandhi doesn't want the consequence of people being murdered.

Stability of goals in self-modifying agents:

- Gandhi doesn't want to kill people.
- Offer Gandhi a pill that will make him enjoy killing people?

- If Gandhi correctly assesses this is what the pill does, Gandhi will refuse the pill, because the current Gandhi doesn't want the consequence of people being murdered.

- Argues for folk theorem that *in general,* rational agents will *preserve their utility functions during self-optimization.*

(Steve Omohundro, 2008. "The Basic AI Drives." In *Proceedings of the First AGI Conference,* eds. Pei Wang and Stan Franklin.)

What do we think we can guess?

- Technologies which *impact upon cognition* will end up mattering most.
- Today's difficulties in constructing AI are due to ignorance of key insights, not inherent mysteriousness.
- It is possible for minds to exist that are *far* more powerful than human.
- An AI over some threshold level of intelligence will recursively self-improve and explode into superintelligence.
- Not all possible agents in mind design space are friendly.
- *If we can obtain certain new insights,* it should be possible to construct a benevolent self-improving Artificial Intelligence.

What do we think we can guess?

- Technologies which *impact upon cognition* will end up mattering most.
- Today's difficulties in constructing AI are due to ignorance of key insights, not inherent mysteriousness.
- It is possible for minds to exist that are *far* more powerful than human.
- An AI over some threshold level of intelligence will recursively self-improve and explode into superintelligence.
- Not all possible agents in mind design space are friendly.
- *If we can obtain certain new insights,* it should be possible to construct a benevolent self-improving Artificial Intelligence.
- And live happily ever after.

More Info:

http://LessWrong.com/
http://Yudkowsky.net/
http://SingInst.org/

Eliezer Yudkowsky
Research Fellow
Singularity Institute for Artificial Intelligence

yudkowsky.net

Yeshiva University
March 2011