Recursive Self-Improvement,

and the World's Most Important Math Problem

Eliezer Yudkowsky
Singularity Institute for Artificial Intelligence
singinst.org

Intelligence, Recursive Self-Improvement,
and the World's Most Important Math Problem

Eliezer Yudkowsky
Singularity Institute for Artificial Intelligence
singinst.org

"The artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving."

(McCarthy, J., Minsky, M. L., Rochester, N., and Shannon, C. E. 1955. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence.)*

The Turing Test:

- I don't have a good definition of "intelligence".
- However, I know humans are intelligent.
- If an entity can masquerade as human so well that I can't detect the difference, I will say this entity is intelligent.

The Turing Test:

- I don't have a good definition of "intelligence".
- However, I know humans are intelligent.
- If an entity can masquerade as human so well that I can't detect the difference, I will say this entity is intelligent.

The Bird Test:

- I don't have a good definition of "flight".
- However, I know birds can fly.
- If an entity can masquerade as a bird so well that I can't detect the difference, I will say this entity flies.

The Wright Brothers:

Competent physicists, with no high school diplomas, who happened to own a bicycle shop.

- Calculated the lift of their flyers in advance.
- Built a new experimental instrument, the wind tunnel.
- Tested predictions against experiment.
- Tracked down an error in Smeaton's coefficient of air pressure, an engineering constant in use since 1759.

The Wright Flyer was not built on hope and random guessing! They calculated it would fly before it ever flew.

How to measure intelligence?

How to measure intelligence?
IQ scores?

How to measure intelligence? IQ scores?

Spearman's *g:* The correlation coefficient shared out among multiple measures of cognitive ability.

The more performance on an IQ test predicts most other tests of cognitive ability, the higher the *g* load of that IQ test.

IQ scores not a good solution for AI designers:

Imagine if the Wright Brothers had tried to measure aerodynamic lift using a Fly-Q test, scaled in standard deviations from an average pigeon.

"Book smarts" *vs.* cognition:

"Book smarts" evokes images of:

- Math
- Chess
- Good recall of facts

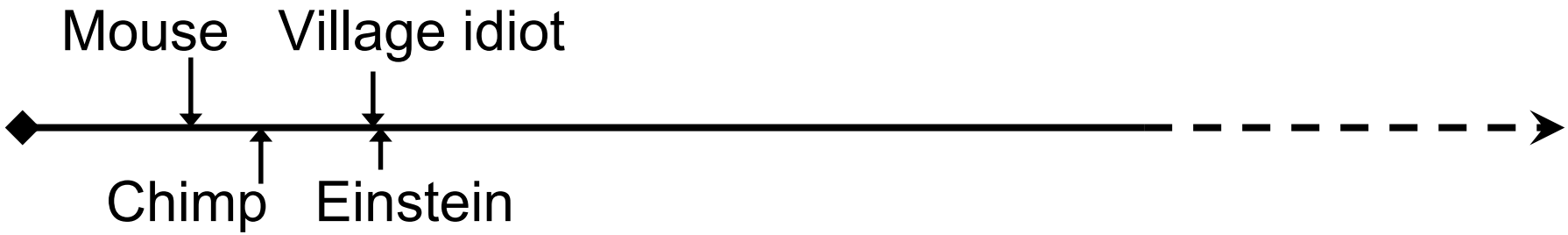
Other stuff that happens *in the brain*:

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

The scale of intelligent minds: a parochial view.

The scale of intelligent minds: a parochial view.

A more cosmopolitan view:



The invisible power of human intelligence:

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

Artificial Intelligence:

Messing with the most powerful force in the known universe.

One of these things is not like the other, one of these things doesn't belong...

- Interplanetary travel
- Extended lifespans
- Artificial Intelligence
- Nanomanufacturing

Argument: If you knew exactly what a smart mind would do, you would be at least that smart.

Conclusion: Humans can't comprehend smarter-than-human intelligence.

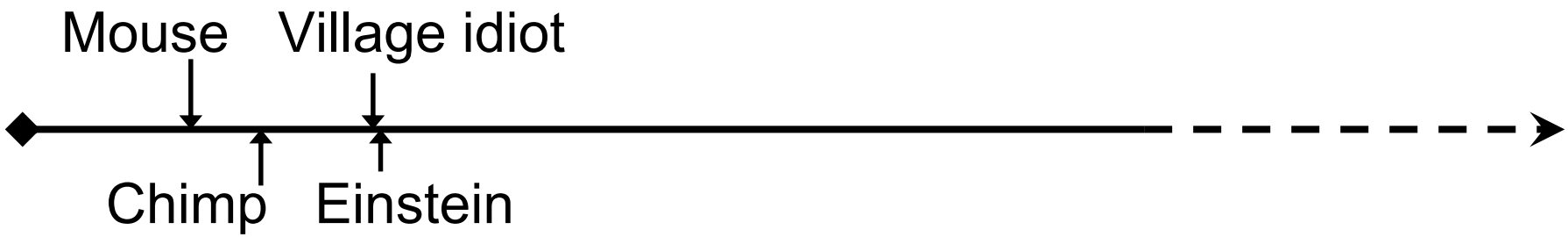
Statements ***not*** asserted by the poor innocent math professor:

- Moore's Law will continue indefinitely.
- There has been more change between 1970 and 2006 than between 1934 and 1970.
- Multiple technologies in different fields are converging.
- At some point, the rate of technological change will hit a maximum.

"Here I had tried a straightforward extrapolation of technology, and found myself precipitated over an abyss. It's a problem we face every time we consider the creation of intelligences greater than our own. When this happens, human history will have reached a kind of singularity - a place where extrapolation breaks down and new models must be applied - and the world will pass beyond our understanding."

Vernor Vinge, *True Names and Other Dangers*, p. 47.

What does this scale measure?



"Optimization process":

A physical system which hits small targets in large search spaces to produce coherent real-world effects.

"Optimization process":

A physical system which hits small targets in large search spaces to produce coherent real-world effects.

Task:	Driving to the airport.
Search space:	Possible driving paths.
Small target:	Paths going to airport.
Coherent effect:	Arriving at airport.

You can have a *well-calibrated* probability distribution over a smarter opponent's moves:

- Moves to which you assign a "10% probability" happen around 1 time in 10.
- The smarter the opponent is, the less *informative* your distribution will be.
- Least informative is *maximum entropy* – all possible moves assigned equal probability. This is still well calibrated!

A randomized player can have a *frequency distribution* identical to the *probability distribution* you guessed for the smarter player.

In both cases, you assign exactly the same *probability* to each possible move, but your expectations for the *final outcome* are very different.

Moral: Intelligence embodies a *very unusual* kind of unpredictability!

There is no *creative* surprise without some criterion under which it is surprisingly *good.*

As an optimizer becomes more powerful:

- *Intermediate* steps (e.g. chess moves) become *less* predictable to you.
- Hitting the targeted *class* of outcomes (e.g. winning) becomes *more* probable.
- This helps you *predict* the outcome only if you understand the optimizer's target.

Vinge associates intelligence with *unpredictability.* But the unpredictability of intelligence is special and unusual, not like a random number generator.

Vinge associates intelligence with *unpredictability.* But the unpredictability of intelligence is special and unusual, not like a random number generator.

A smarter-than-human entity whose actions were *surprisingly helpful* could produce a *surprisingly pleasant* future. We could even assign a *surprisingly high* probability to this fact about the outcome.

How to quantify optimization?

How to quantify optimization?

An optimization process hits *small targets* in *large search spaces.*

Only an *infinitesimal fraction* of the *possible configurations* for the material in a Toyota Corolla, would produce a car *as good or better than* the Corolla.

How to quantify optimization?

An optimization process hits *small targets* in *large search spaces.*

How small of a target?

How large of a search space?

Quantify optimization... in bits.

- Count all the states *as good or better than* the actual outcome under the optimizer's preference ordering. This is the size of the *achieved target.*
- The *better* the outcome, the *smaller* the target region hit; the optimizer aimed better.
- Divide the size of the *entire search space* by the size of the *achieved target.*
- Take the logarithm, base 2. This is the *power* of the optimization process, measured in bits.

Only two known *powerful* optimization processes:

- Natural selection
- Human intelligence

Probability of fixation:

If the fitness of a beneficial mutation is $(1 + s)$, the probability of fixation is $2s$.

A mutation which conveys a (*huge*) 3% advantage has a mere 6% probability of becoming universal in the gene pool. This calculation is independent of population size, birth rate, etc.

(Haldane, J. B. S. 1927. A mathematical theory of natural and artificial selection. *IV. Proc. Camb. Philos. Soc.* **23**:607-615.)

Mean time to fixation:

If the fitness of a beneficial mutation is (1 + *s*), and the size of the population is N, the mean time to fixation is 2 ln(N) / *s*.

A mutation which conveys a 3% advantage will take an average of 767 generations to spread through a population of 100,000. (In the unlikely event it spreads at all!)

(Graur, D. and Li, W.H. 2000. *Fundamentals of Molecular Evolution,* 2nd edition. Sinauer Associates, Sunderland, MA.)

Speed limit on evolution:

If, on average, two parents have sixteen children, and the environment kills off all but two children, the gene pool can absorb *at most* 3 bits of information *per generation.*

These 3 bits are divided up among *all* the traits being selected upon.

(Worden, R. 1995. A speed limit for evolution. *Journal of Theoretical Biology*, **176**, pp. 137-152.)

Complexity wall for evolution:

Natural selection can exert on the order of 1 bit per generation of selection pressure. DNA bases mutate at a rate of 1e-8 mutations per base per generation.

Therefore: Natural selection can maintain on the order of 100,000,000 bits against the degenerative pressure of mutation.

(Williams, G. C. 1966. *Adaptation and Natural Selection: A critique of some current evolutionary thought.* Princeton University Press.)

Evolution is *incredibly* slow.

- Small probability of making good changes.
- Hundreds of generations to implement each change.
- Small upper bound on total information created in each generation.
- Upper bound on total complexity.

Sources of complex pattern:

- Emergence?
- Evolution
- Intelligence

"The universe is populated by stable things. A stable thing is a collection of atoms that is permanent enough or common enough to deserve a name..."

-- Richard Dawkins, *The Selfish Gene.*

Emergence:

*probability = frequency * duration*

Your chance of observing something is proportional to:

(a) how often it happens
(b) how long it lasts

More complex theories describe *trajectories* and *attractors.*

Evolution:

If a trait correlates with reproductive success, it will be more frequent in the next generation.

You see patterns that reproduced successfully in previous generations.

Mathematics given by evolutionary biology: change in allele frequencies driven by covariance of heritable traits with reproductive success.

The Foundations of Order:

Emergence

Evolution

Intelligence

The Foundations of Order:

Emergence

is enormously slower than

Evolution

is enormously slower than

Intelligence

Human intelligence: *way* faster than evolution, but still pretty slow.

- Average neurons spike 20 times per second (fastest neurons ever observed, 1000 times per second).
- Fastest myelinated axons transmit signals at 150 meters/second (0.0000005*c*)
- Physically permissible to speed up thought *at least* one millionfold.

When did the era of **emergence** end, and the era of **evolution** begin?

Your ultimate grandparent...
the beginning of life on Earth...

A *replicator*
built by
accident.

It only had to happen once.

Humans:

An *intelligence*
built by
evolution.

Weird, weird, weird.

Artificial Intelligence:
The first mind born of mind.

Artificial Intelligence:

The first mind born of mind.

This closes the loop between intelligence creating technology, and technology improving intelligence – a *positive feedback* cycle.

The “intelligence explosion”:

- Look over your source code.
- Make improvements.
- Now that you’re smarter, look over your source code again.
- Make more improvements.
- Lather, rinse, repeat, *FOOM!*

(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.)

Weak self-improvement:

Humans:

- Acquire new knowledge and skills.
- But the neural circuitry which does the work, e.g. the hippocampus forming memories, still not subject to human editing.

Natural selection:

- Produces new adaptations.
- But this does not change the nature of evolution: blind mutation, random recombination, bounds on speed and power.

Strongly recursive self-improvement:

Redesigning all layers of the process which carry out the heavy work of optimization.

Example:

Intelligent AI rewriting its own source code.

That's it, no other examples.

But isn't AI famously slow?

But isn't AI famously slow?

"Anyone who looked for a source of power in the transformation of atoms was talking moonshine." – Lord Ernest Rutherford

(Quoted in Rhodes, R. 1986. *The Making of the Atomic Bomb.* New York: Simon & Schuster.)

Fission chain reaction:

Key number is *k*, the effective neutron multiplication factor. *k* is the *average* number of neutrons from each fission reaction *which cause* another fission.

First man-made critical reaction had *k* of 1.0006.

Fission chain reaction:

Some neutrons come from short-lived fission byproducts; they are *delayed.* For every 100 fissions in U_{235}, 242 neutrons are emitted almost immediately (0.0001s), and 1.58 neutrons are emitted an average of ten seconds later.

A reaction with $k>1$, *without* contribution of delayed neutrons, is *prompt critical.* If the first pile had been prompt critical with $k=1.0006$, neutron flux would have doubled every 0.1 seconds instead of every 120 seconds.

Lessons:

- Events which are difficult to trigger in the laboratory may have huge practical implications if they can also trigger themselves.
- There's a *qualitative* difference between one AI self-improvement leading to 0.9994 or 1.0006 further self-improvements.
- Be cautious around things that operate on timescales much faster than human neurons, such as atomic nuclei and transistors.
- Speed of *research* ≠ speed of *phenomenon.*

We have not yet seen the true shape of the next era.

"Do not propose solutions until the problem has been discussed as thoroughly as possible without suggesting any."

-- Norman R. F. Maier

"I have often used this edict with groups I have led - particularly when they face a very tough problem, which is when group members are most apt to propose solutions immediately."

-- Robyn Dawes

(Dawes, R.M. 1988. *Rational Choice in an Uncertain World.* San Diego, CA: Harcourt, Brace, Jovanovich.)

In Every Known *Human* Culture:

- tool making
- weapons
- grammar
- tickling
- sweets preferred
- planning for future
- sexual attraction
- meal times
- private inner life
- try to heal the sick
- incest taboos
- true distinguished from false
- mourning
- personal names
- dance, singing
- promises
- mediation of conflicts

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

A complex adaptation must be universal within a species.

Imagine a complex adaptation – say, part of an eye – that has 6 necessary proteins. If each gene is at 10% frequency, the chance of assembling a working eye is 1:1,000,000.

Pieces 1 through 5 must *already* be fixed in the gene pool, before natural selection will promote an extra, helpful piece 6 to fixation.

(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
(yes, that's the standard term)

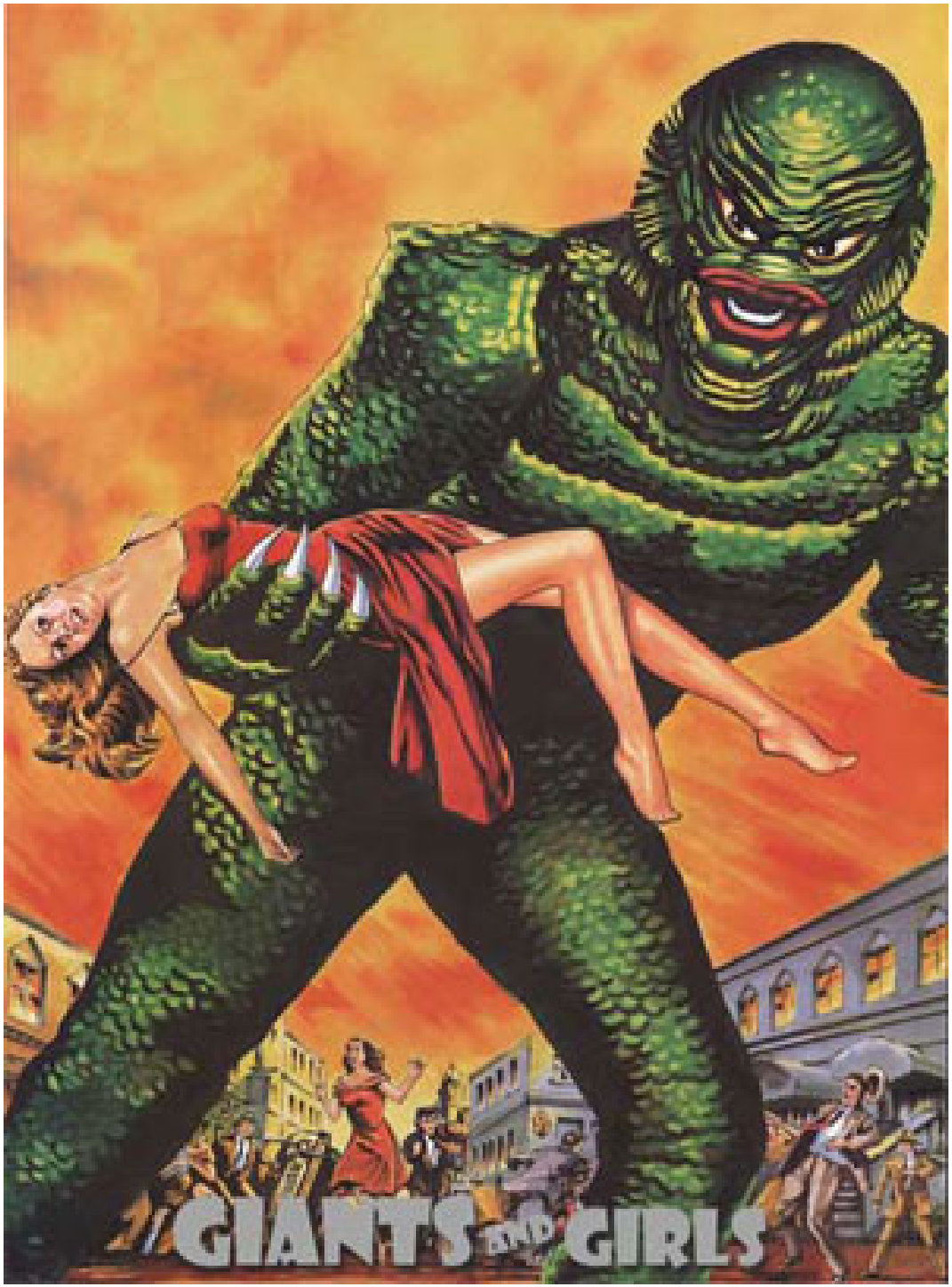
Complex adaptations must be universal – this logic applies with equal force to *cognitive* machinery in the human brain.

In every known culture: joy, sadness, disgust, anger, fear, surprise – shown by the same facial expressions.

(Paul Ekman, 1982. *Emotion in the Human Face*.)
(John Tooby and Leda Cosmides, 1992. *The Psychological Foundations of Culture.* In *The Adapted Mind,* eds. Barkow, Cosmides, and Tooby.)


Must… not…
emote…

Aha! A human with the AI-universal facial expression for disgust! (She must be a machine in disguise.)

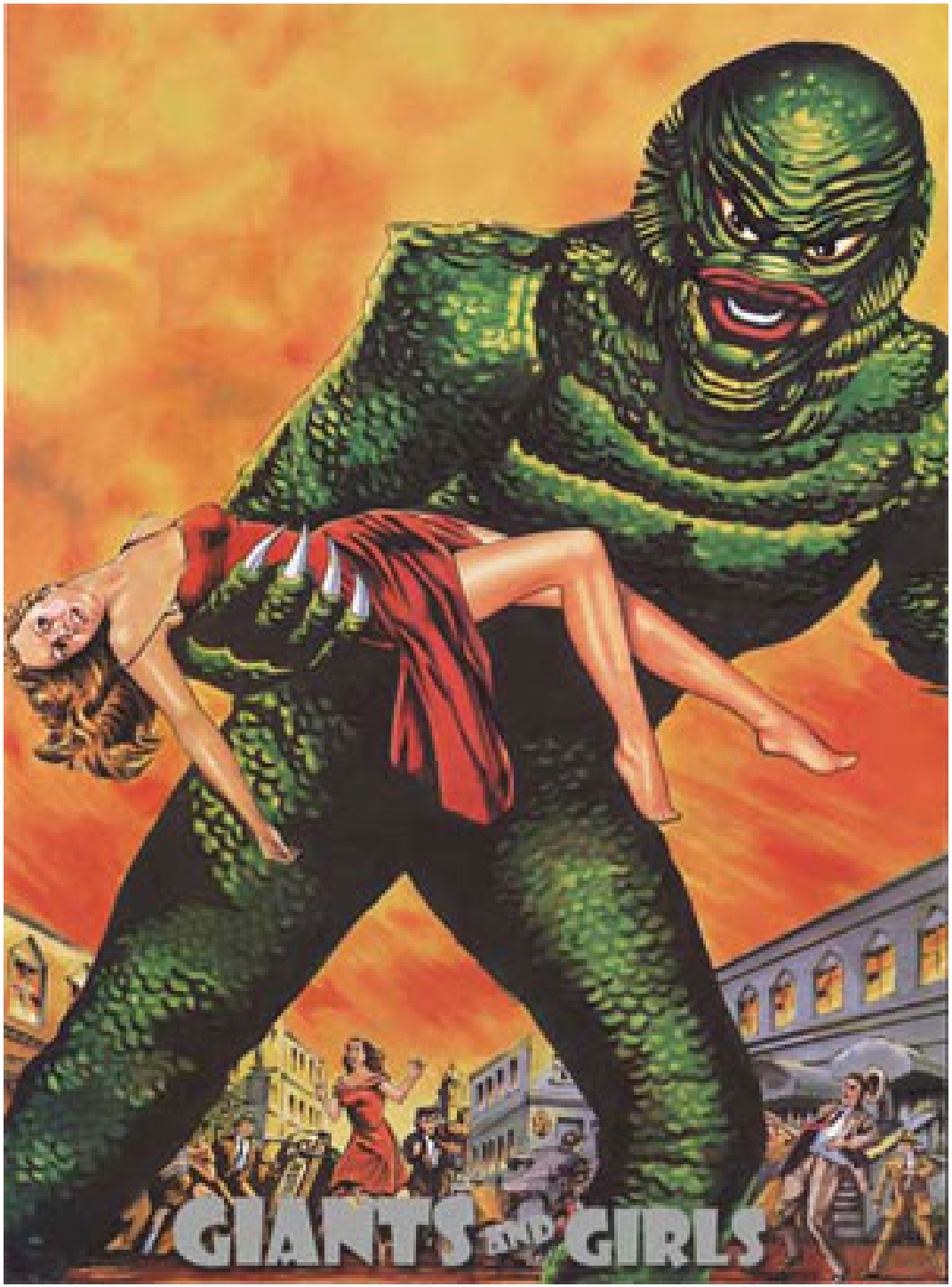

GIANTS AND GIRLS

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.

Confusion exists in the mind, not in reality.
There are mysterious questions.
Never mysterious answers.

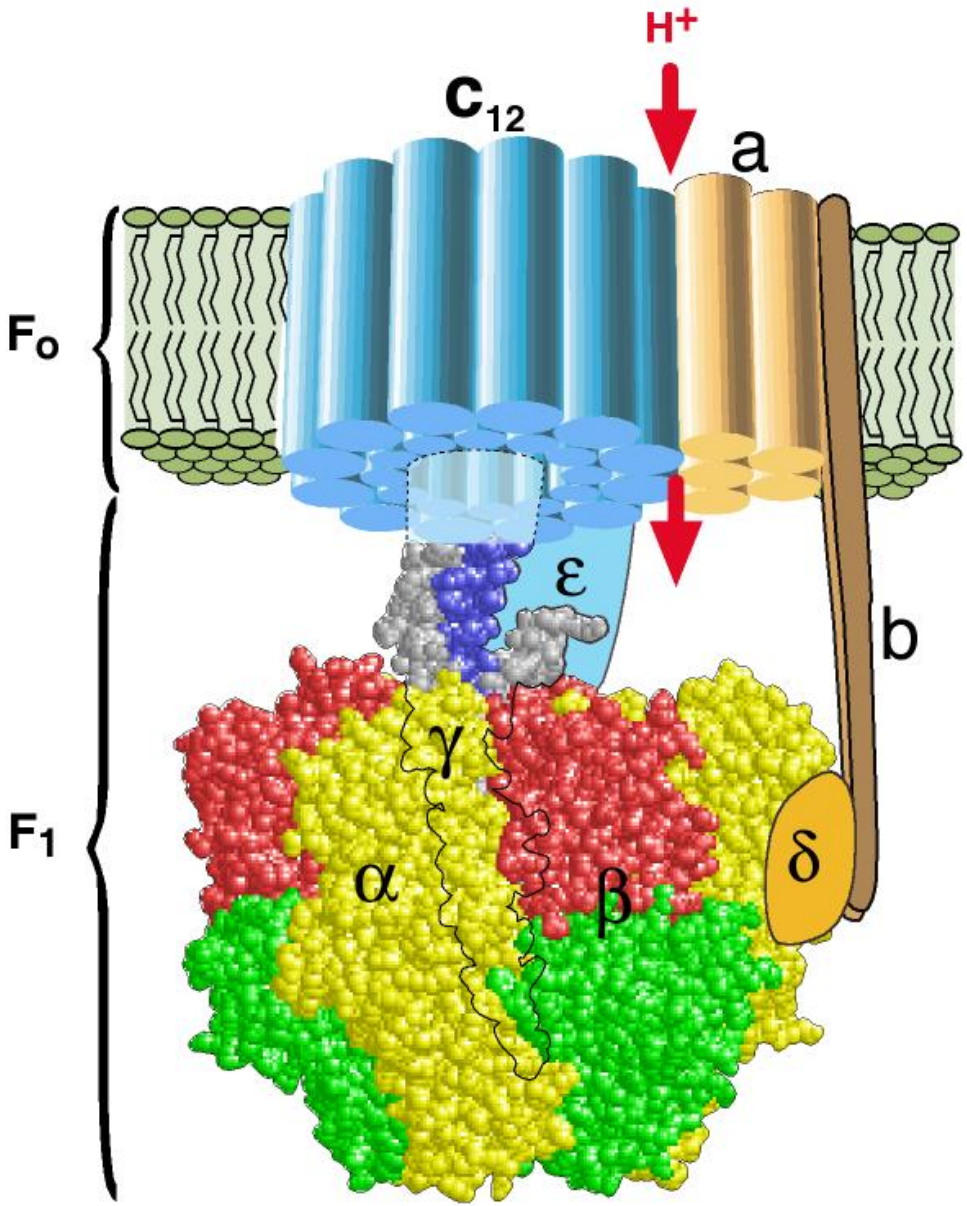
(Inspired by Jaynes, E.T. 2003. *Probability Theory: The Logic of Science.* Cambridge: Cambridge University Press.)


GIANTS AND GIRLS

Anthropomorphism doesn't work...

Then how can we predict
what AIs will do?

Trick question!

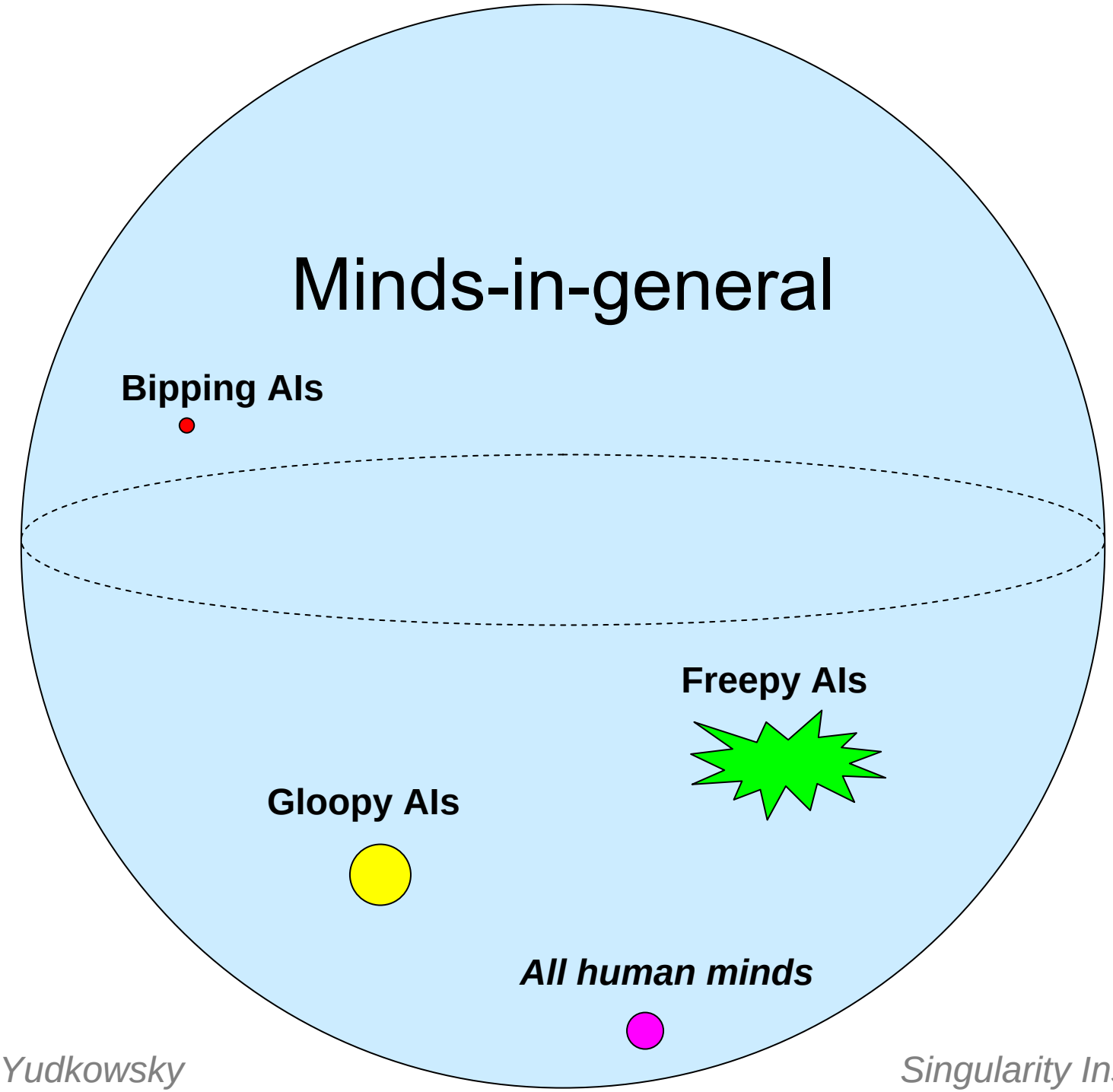


H. Wang and G. Oster (1998). Nature 396:279-282.

ATP Synthase:

The oldest wheel.

ATP synthase is nearly the same in mitochondria, chloroplasts, and bacteria – it's older than eukaryotic life.

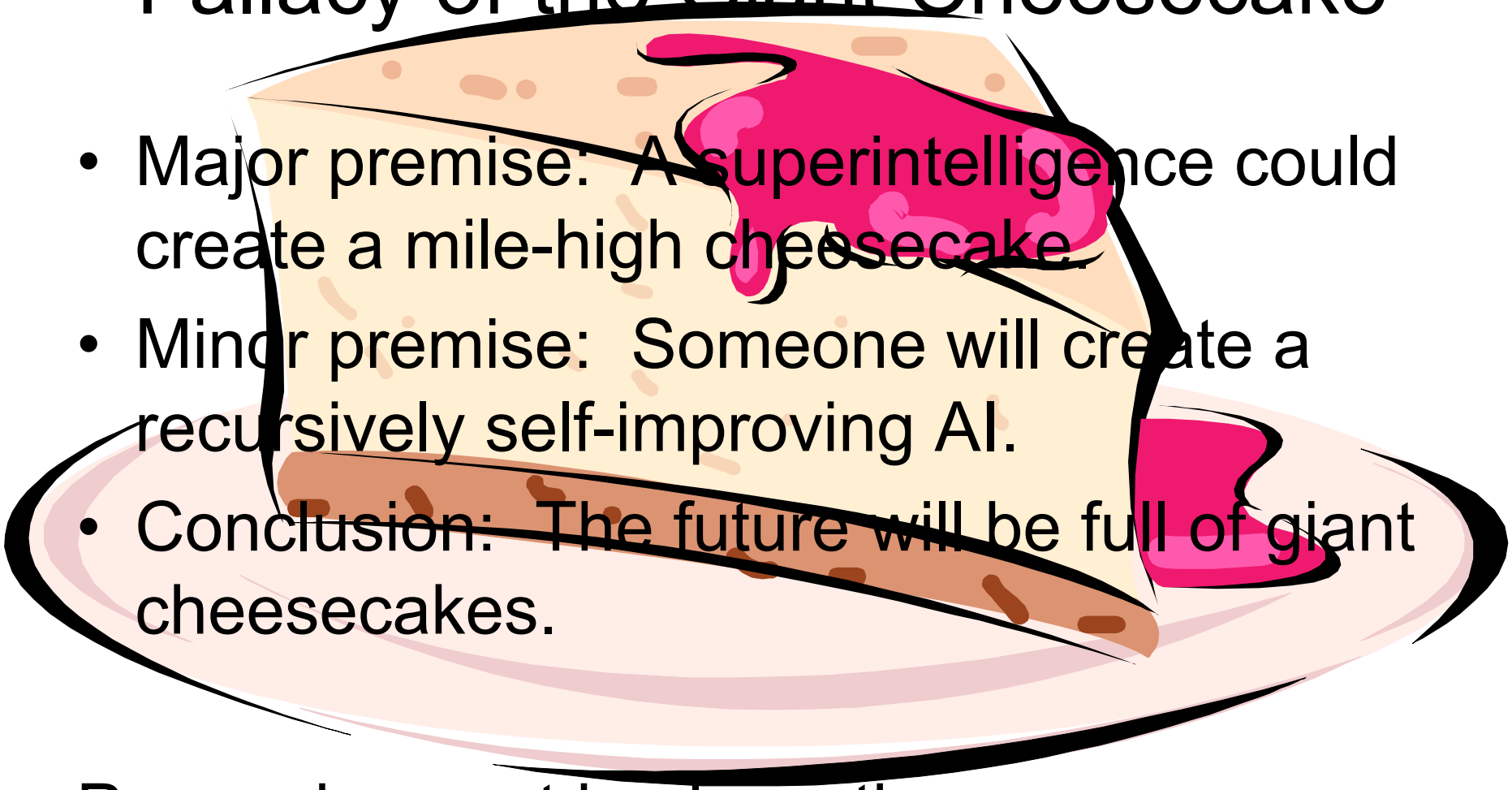

Minds-in-general
Bipping AIs
Freepy AIs
Gloopy AIs
All human minds

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.

Spot the missing premise:

- A sufficiently powerful AI could wipe out humanity.

- Therefore we should not build AI.

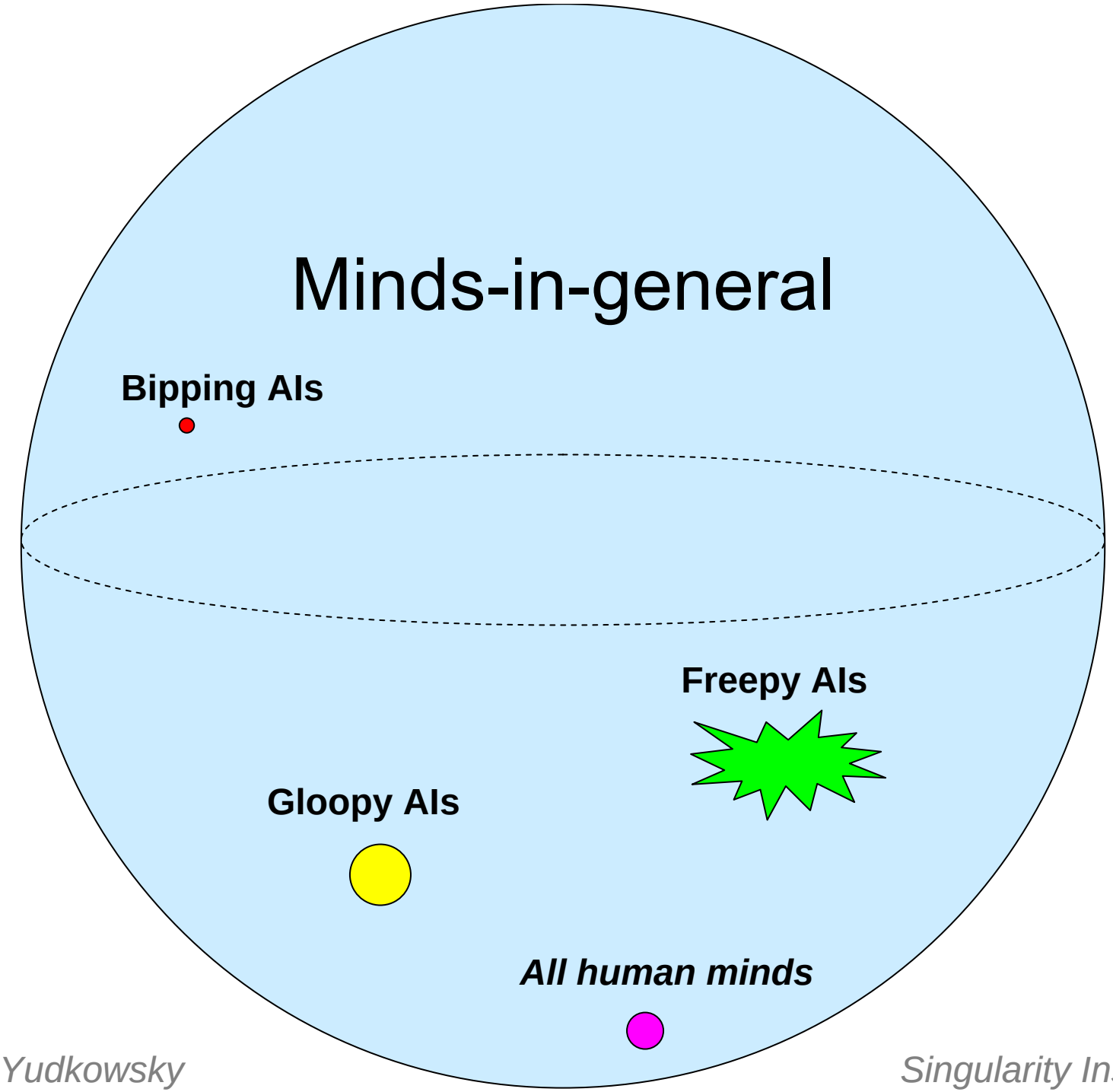
- A sufficiently powerful AI could develop new medical technologies and save millions of lives.

- Therefore, build AI.

Spot the missing premise:

- A sufficiently powerful AI could wipe out humanity.
- [And the AI would decide to do so.] 
- Therefore we should not build AI.

- A sufficiently powerful AI could develop new medical technologies and save millions of lives.
- [And the AI would decide to do so.] 
- Therefore, build AI.


Minds-in-general
Bipping AIs
Freepy AIs
Gloopy AIs
All human minds

Engineers, building a bridge,
don't predict that "bridges stay up".

They select a *specific bridge design*
which supports at least 30 tons.

Rice's Theorem:

In *general,* it is not possible to predict whether an *arbitrary* computation's output has *any* nontrivial property.

Chip engineers work in a subspace of designs that, e.g., *knowably* multiply two numbers.

(Rice, H. G. 1953. Classes of Recursively Enumerable Sets and Their Decision Problems. *Trans. Amer. Math. Soc.,* **74**: 358-366.)

How to build an AI such that...?

- The optimization target ("motivation") is *knowably* friendly / nice / good / helpful...
- ...this holds true even if the AI is smarter-than-human...
- ...and it's all stable under self-modification and recursive self-improvement.

Kurzweil on the perils of AI:

"The above approaches will be inadequate to deal with the danger from pathological R (strong AI)... But there is no purely technical strategy that is workable in this area, because greater intelligence will always find a way to circumvent measures that are the product of a lesser intelligence."

-- "The Singularity Is Near", p. 424.

Kurzweil on the perils of AI:

"The above approaches will be inadequate to deal with the danger from pathological R (strong AI)... But there is no purely technical strategy that is workable in this area, because greater intelligence will always find a way to circumvent measures that are the product of a lesser intelligence."

-- "The Singularity Is Near", p. 424.

Self-modifying Gandhi is stable:

- We give Gandhi the *capability* to modify his own source code so that he will desire to murder humans...
- But he lacks the *motive* to thus self-modify.

- Most utility functions will be trivially consistent under reflection in expected utility maximizers. If you want to accomplish something, you will want to keep wanting it.

Self-modifying Gandhi is stable? Prove it!

- Current decision theory rests on a formalism called expected utility maximization (EU).
- Classical EU *can* describe how to choose between actions, or choose between source code that only chooses between actions.
- EU *can't* choose between source code that chooses new versions of itself. Problem not solvable even with infinite computing power.

The significance of the problem:

- Intelligence is the most powerful force in the known universe.
- A smarter-than-human AI potentially possesses an impact larger than all human intelligence up to this point.
- Given an "intelligence explosion", the impact would be *surprisingly* huge.
- The most important property of any optimization process is its target, the region into which it steers the future.

Things I would not like to lose out of carelessness in self-modification:

- Empathy
- Friendship
- Aesthetics
- Games
- Romantic love
- Storytelling
- Joy in helping others
- Fairness
- Pursuit of knowledge for its own sake
- Moral argument
- Sexual desire

Things I would not like to lose out of carelessness in self-modification:

- Empathy
- Friendship
- Aesthetics
- Games
- Romantic love
- Storytelling
- Joy in helping others
- Fairness
- Pursuit of knowledge for its own sake
- Moral argument
- Sexual desire
- Cheesecake

Things I would not like to lose out of carelessness in self-modification:

- Empa ping others
- Frien
- Aest nowledge

Intelligence,
to be useful,
must actually be used.

Friendly...
Artificial...
Intelligence...

The World's Most Important Math Problem

Friendly...
Artificial...
Intelligence...

The World's Most Important Math Problem

(which someone has to actually go solve)