

# What can evolution tell us about the feasibility of artificial intelligence?

Carl Shulman

Singularity Institute for Artificial Intelligence

# Artificial intelligence

Systems that can learn to perform almost any tasks that humans can, including scientific, economic, and military tasks

Robust substitutes for human labor

Created in a computer, not the bedroom or a test tube



# Why AI matters

Software can be copied, run faster with faster hardware, easily edited

Superabundant skilled labor: extreme prosperity, cures for all diseases, &c.

Risk of instability, conflict, human extinction (Hawking 1998, Bostrom 2002)



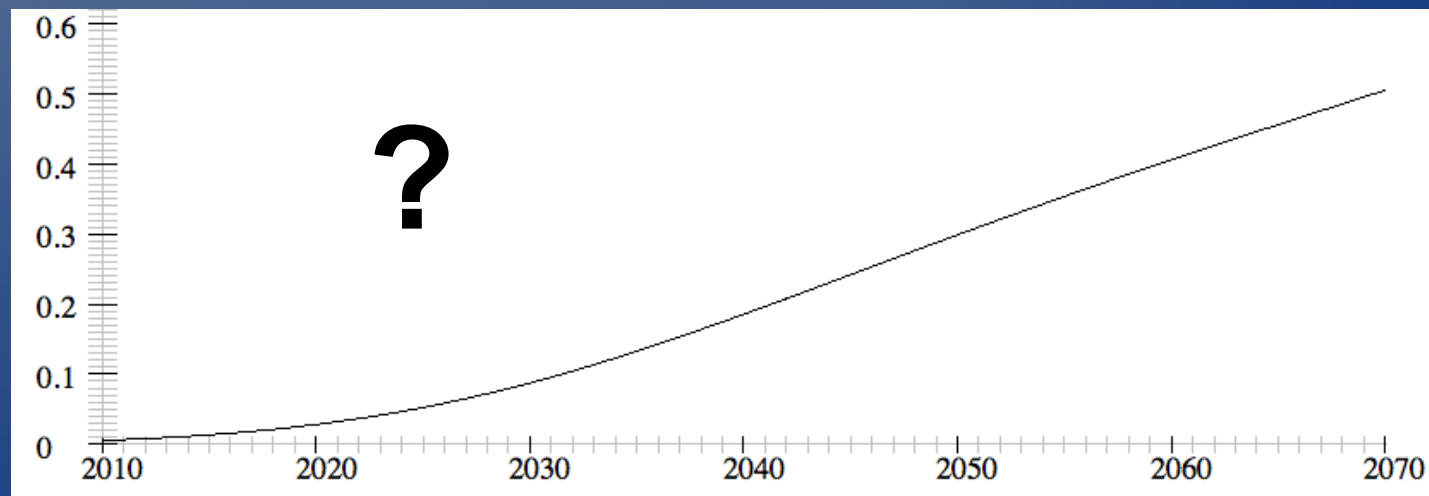
# AI timelines matter

Probability of AI:

... in 2040? (*The Singularity is Near*)

... in 2100? (impact on climate change)

... in 3000? (mostly theoretical interest)



# Two kinds of evidence

How good are we at solving AI problems?

Past progress in machine learning algorithms,  
computational neuroscience, hardware

How numerous and how hard are the problems to  
be solved?

One example of the development of human  
intelligence (biological evolution on Earth)

# Two kinds of evidence

How good are we at solving AI problems?

Past progress in machine learning algorithms,  
computational neuroscience, hardware

**How numerous and how hard are the  
problems to be solved?**

**One example of the development of human  
intelligence (biological evolution on Earth)**

# Hard for evolution, hard for us?

Some capacities are relatively hard for evolution to create but relatively easy for 21<sup>st</sup> century human civilizations

Fire, supersonic flight, nuclear fission

Evolution: large populations, many generations allow random mutation to explore nearby improvements that enhance fitness

Humans: theory, planning, non-random search, foresight; less time and fewer resources



# Evolution as upper bound

Knowing that evolution can produce a feature  
give an upper bound for its difficulty

If we expect a big speedup from intelligent  
search, difficulty for us should be well below the  
evolutionary bound

Speedup may vary by problem, but still evidence

So it matters how hard it is to evolve features of  
human intelligence

# I. Naïve estimates untrustworthy

# Intuitively, not too terribly hard

Human-level intelligence evolved on Earth

Sonar, photosynthesis, flight, *etc.* also evolved on Earth

So perhaps evolving intelligence and evolving sonar are both about what you'd expect from 4 billion years of evolution on a typical life-bearing planet

# Intuitively, not too terribly hard

Human-level intelligence evolved on Earth

Sonar, photosynthesis, flight, *etc.* also evolved on Earth

A naïve engineer estimating from evolution might suppose sonar and AI are comparably difficult

# Observer selection asymmetry

It probably isn't  
ridiculously  
hard to evolve sonar.



## SENSIBLE

If sonar was hard to evolve, we  
would not expect to observe it

It probably isn't  
ridiculously  
hard to evolve  
intelligence.



## SUSPECT

Intelligent creatures will *always*  
observe a world in which  
intelligence has evolved

# Earth is typical of ... ?



**PLANETS**

**naïve view**



**PLANETS WITH  
CIVILIZATION**

**one anthropic view**

# Self-Sampling Assumption

“One should reason as if one were a random sample from the set of all observers in one’s reference class.” (Bostrom 2002)

Leaves open choice of reference class

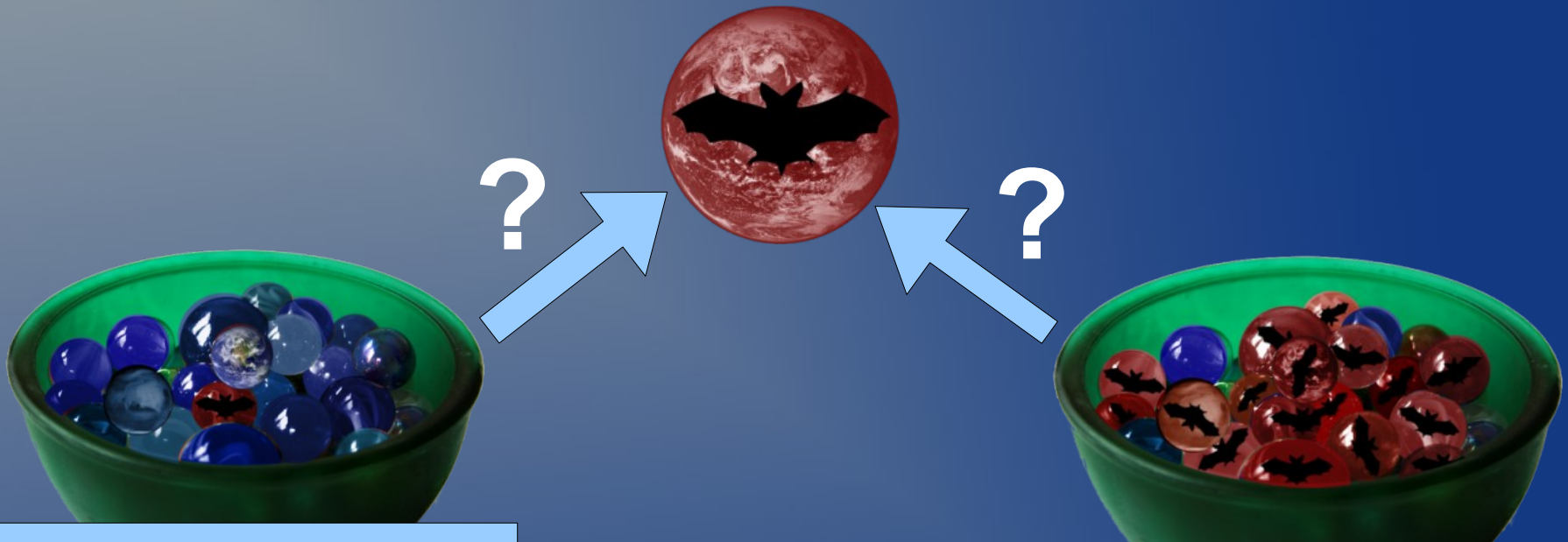
Alternative assumptions also plausible (e.g., self-indication)

# SSA-civs

“SSA-civs”: reason as if your planet is a random sample of the reference class of planets with civilizations



# SSA enables calculations



**Theory 1: 1% of planets have biosonar**

**Theory 2: 90% of planets have biosonar**

$$\frac{P(\text{sonar}|\text{T2})P(\text{T2})}{P(\text{sonar}|\text{T2})P(\text{T2}) + P(\text{sonar}|\text{T1})P(\text{T1})} = \frac{(.900)(.500)}{(.900)(.500) + (.010)(.500)} \approx 0.989$$

**II.**

**A simple example:  
reasoning from SSA-civs**

# Applying SSA-civs



First piece of data: when humans evolved on Earth

Earth: 4.6 billion years old, earliest fossils of life 3.8 billion, Cambrian explosion 580 million, mammals 280 million, primates 55–85 million, hominids 14–18 million, humans ~1 million

Habitable for another ~1.1 billion years

# What would early intelligence have told us?

Imagine if intelligence had arisen in the first 1% of Earth's habitable period, instead of 4/5ths of the way through

Highly likely if evolution of intelligence is very easy,  
very unlikely if hard

Intelligence isn't *very* easy to evolve

# Hard steps all look alike

Suppose we need to pick five locks by trial-and-error

Ordinarily, can infer lock difficulty from typical opening times

But if all locks must be picked, in a row, in an hour, and you only hear about successes ...

After a (low!) threshold, opening time gives little info about lock difficulty

Evolutionary “locks”: abiogenesis? Sex? Intelligence?

<i>If Done in 1</i>	<i>Five Steps</i>					<i>Left</i>
Difficulty	.01	.1	1	10	100	–
Average	.0096	.0745	.2021	.2366	.2372	.2340
Deviation	.0096	.0722	.1643	.1825	.1830	.1820

Table 1: Simulation of Five Steps with x10 Difficulty Increments

# Hard steps and timing

With a single hard step, no strong prediction about timing

With multiple hard steps, similar intervals between hard steps

Interval from last hard step to end of habitable period similar to interval between hard steps

1.1 billion years of habitability left; such an interval is unlikely if there are more than 7 hard steps



# Hard steps near humanity?

If one is confident particular steps are hard (abiogenesis, multicellularity, brains), few left for near humanity

Unlikely more than one hard step since the emergence of primates

Some hard steps may have occurred early but not been noticeable (neural architecture scaling well)



# Hard steps near humanity?

Import for AI:

Early hard steps appear less relevant to AI design

Don't need to design multicellularity, already have computers

Animal nervous systems easier to study, dissect

Perhaps little warning from chimp- or mouse-level AI before human-level AI



# What does SSA-civs plus timing tell us?

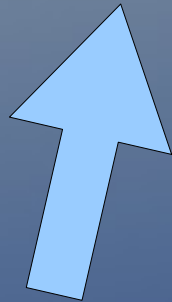
Evolution of intelligence not very easy

Probably 0 to 7 sequential, stochastic hard steps

Probably no more than one hard step since primates



# Applying SSA-civs



Second piece of data:  
convergent evolution of  
intelligence

Humans arose in a world that  
has:

Chimpanzees

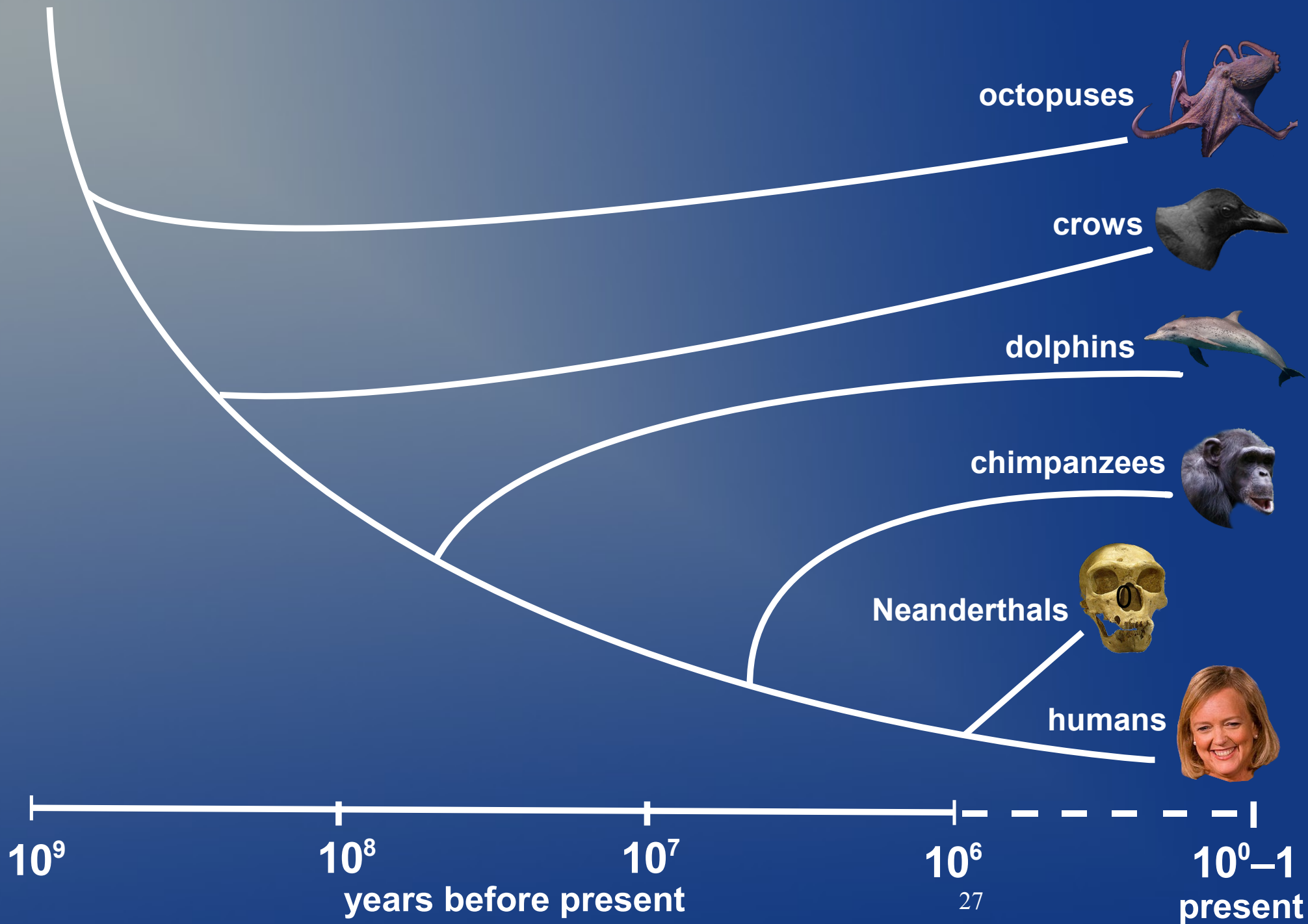
Dolphins

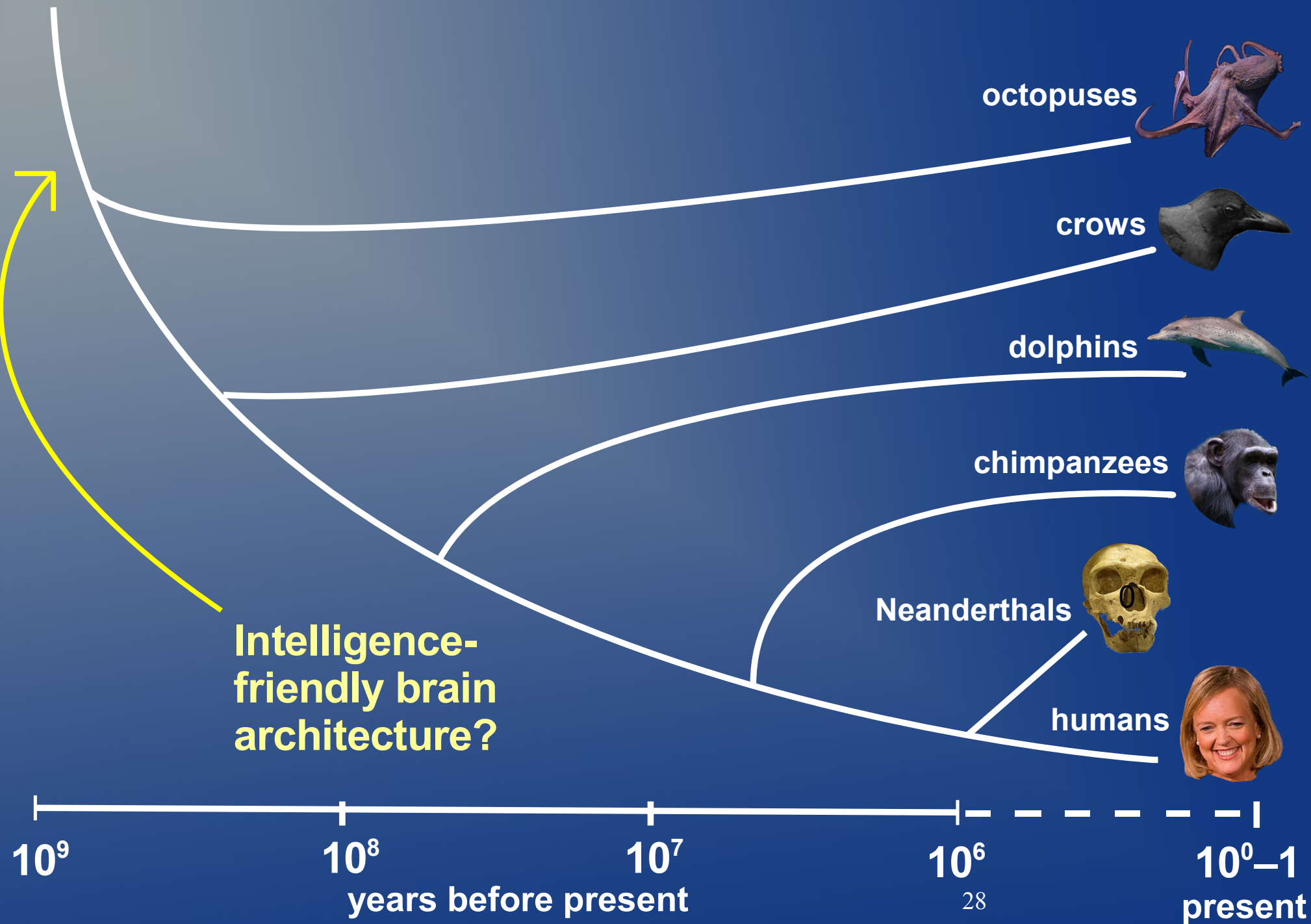
Crows

Octopuses

# Convergent evolution—smart animals







# Convergent evolution of intelligence

Fairly impressive animal intelligence seems to have evolved in several branches of vertebrates, and even some invertebrates (e.g., octopuses)

Rules out a hard step much after the Cambrian to produce that level of intelligence (but nervous systems have shared ancient origin)

# What does SSA-civs plus animal intelligence tell us?

Octopus, crow, elephant intelligence relatively easy to evolve from flatworm-like common ancestor, so either:

Human-level intelligence evolves relatively easily (does not require observer-selection effect after non-nervous system ancestor)

Human intelligence is a hard step from chimp-, crow-, octopus-level intelligence; or

The brain architecture of the common ancestor is selected via anthropic effects to be easily extensible



# **III. Alternative anthropic principles**

# Earth is typical of ... ?



**PLANETS**

**naïve view**



**... ?**

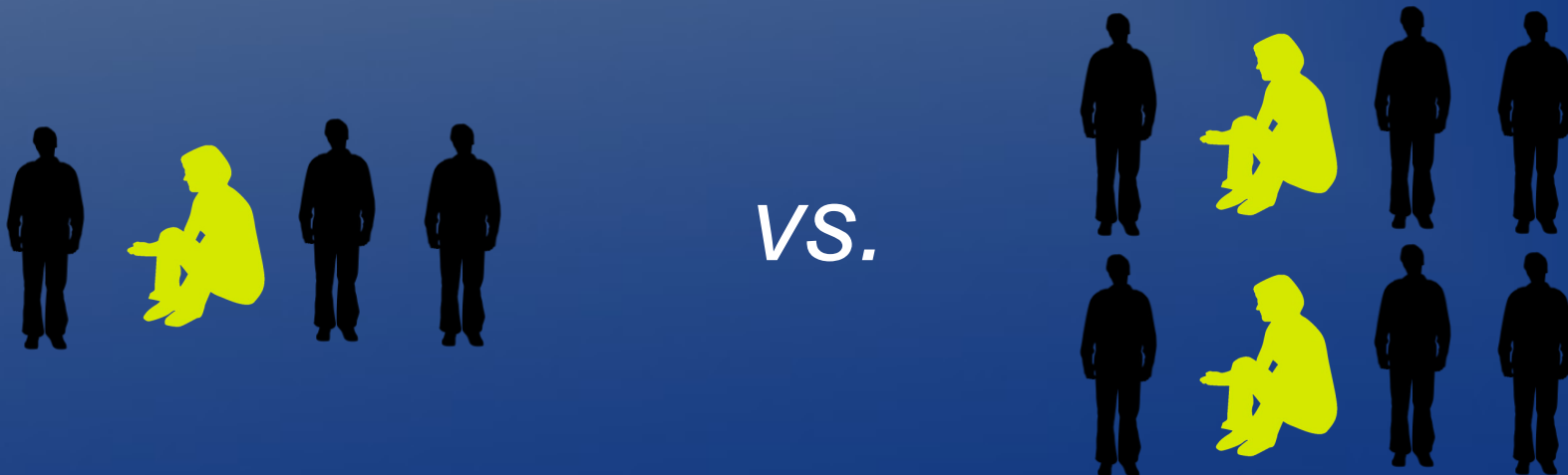
**which anthropic  
view?**

# Two dimensions that might lead us to prefer certain hypotheses

Relative frequency of your experiences



Absolute number of your experiences



# SSA is about proportions

Self-Sampling Assumption favors hypotheses that predict greater relative proportions of your experiences

Leaves open choice of reference class

SSA-civs was a modification of SSA that chose civilizations instead of observers

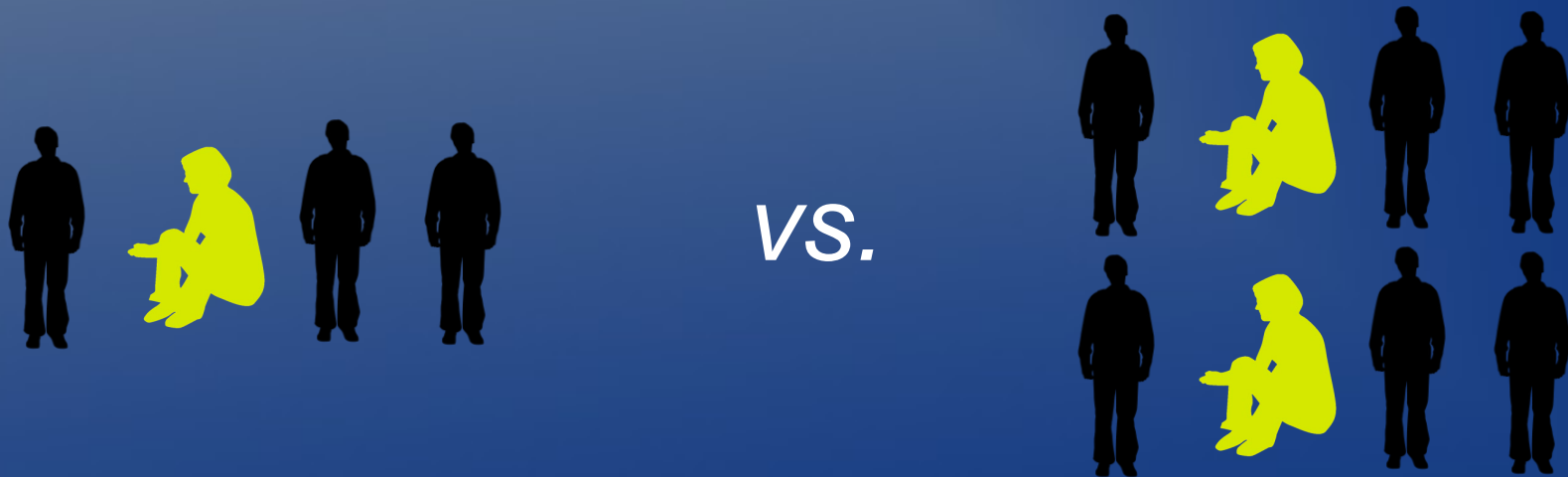


# SIA is about absolute numbers

Self-Indication Assumption favors hypotheses that predict greater relative proportions of your experiences

Leaves open choice of reference class

SSA-civs was a modification of SSA that chose civilizations instead of observers



# Experience lottery

Radio show flips a coin:

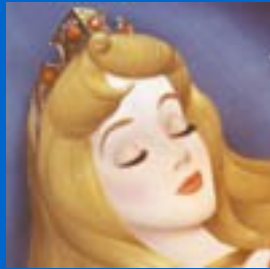
If Heads: calls 1 number at random from directory

If Tails: calls 100 numbers at random from directory

You get a call: how likely is it that the coin came up Tails?

$$\frac{P(\text{call}|\text{T})P(\text{T})}{P(\text{call}|\text{T})P(\text{T}) + P(\text{call}|\text{H})P(\text{H})} \approx 0.99$$

# Sleeping Beauty

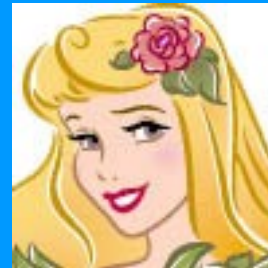
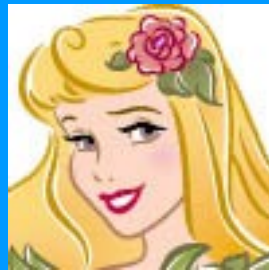


Sunday

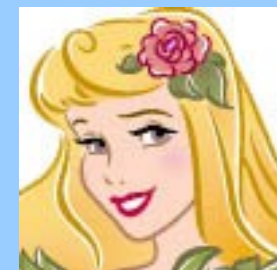
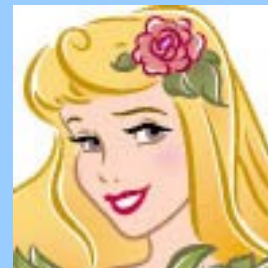
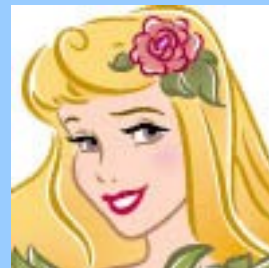
Monday

Tuesday

**HEADS**



**TAILS**



# SIA favors early AI

Favors theories in which your experiences are common

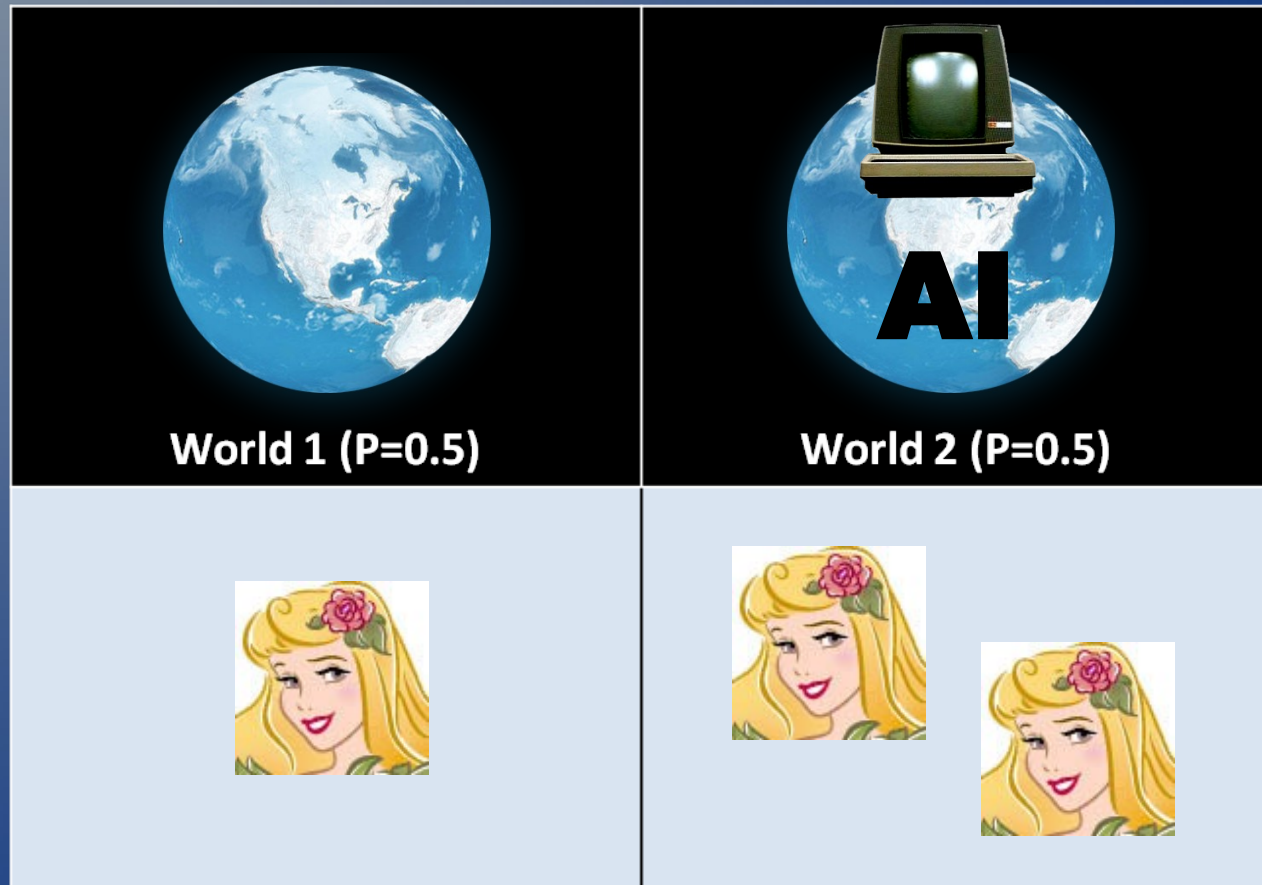
And so, theories in which intelligence is relatively easy to evolve

Fermi paradox (no aliens) and late emergence of intelligence on Earth, somewhat constrains rate of civilization evolution

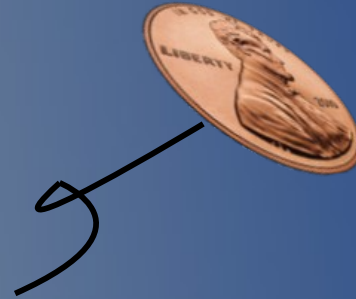
But SIA favors intelligence being as evolvable as possible, subject to our data

And so, theories in which AI is easy to design

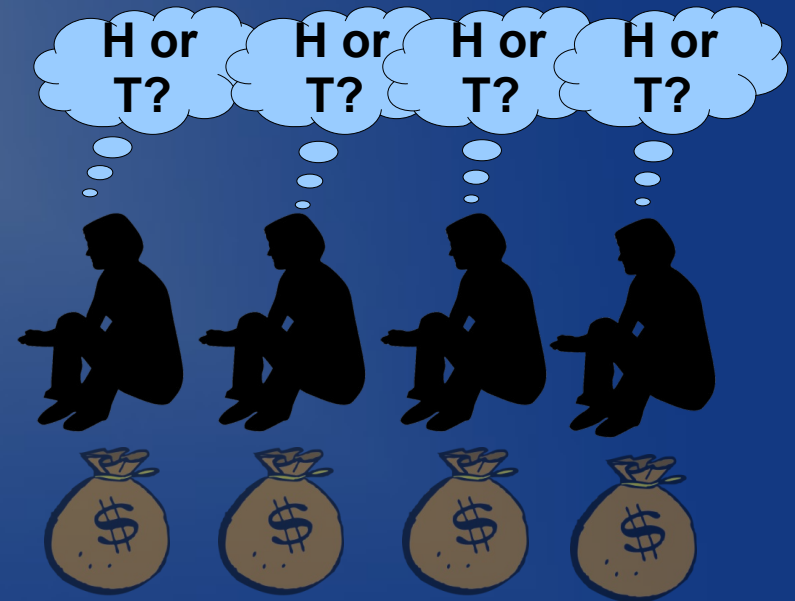
# SIA favors early AI



# One motive: collective betting



**HEADS**



**TAILS**

# IV. Recap and Takeaways

# Inferring intelligence is not like inferring sonar

It probably isn't  
ridiculously  
hard to evolve sonar.



It probably isn't  
ridiculously  
hard to evolve  
intelligence.



# What sort of observer selection principles should we use?



# SSA-civs

Sensitive to initial data from evolution

Data does update our credences about  
evolvability of intelligence

But likelihood ratios are not extreme; answer  
depends on initial credences

# SSA-civs

Intelligence is not *very* easy to evolve.

0–7 hard steps

Major possibilities:



Neurons observer-selected to be extensible

Intelligence is easy (or is easy given flatworm behavior)

There's a hard step between monkeys/crows/octopuses and humans

# SIA

Strong update toward evolution of intelligence  
being easy

Almost an *a priori* argument: likelihood ratios  
overwhelm most priors, almost regardless of  
data

But SIA requires much counterintuitive bullet-  
biting

# Anthropics is relevant to evolution, AI

Institutions concerned about long-term AI forecasting should consider funding anthropics research



Thanks for listening

Reach me at:  
*carl.shulman@post.harvard.edu*