What could we do about intelligence explosion?

Daniel Dewey
daniel.dewey@philosophy.ox.ac.uk

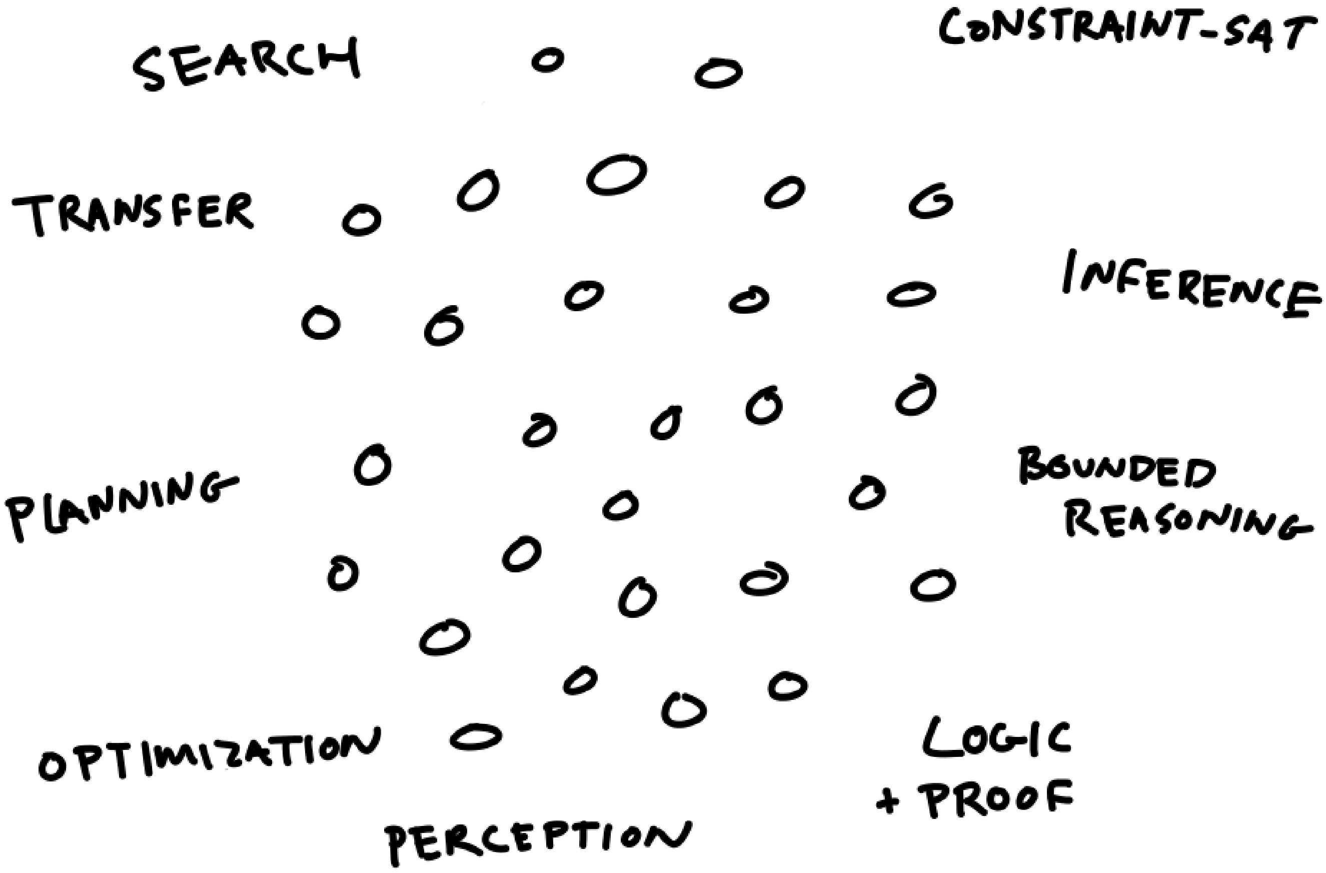
May 2014

Programme on the Impacts of Future Technology
UNIVERSITY OF OXFORD

OXFORD MARTIN SCHOOL

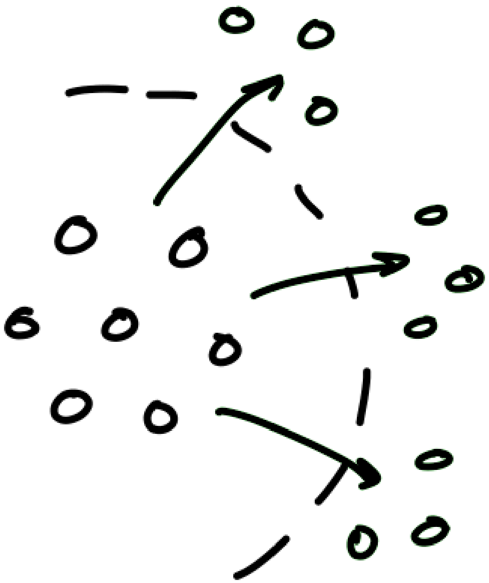
Future of Humanity Institute
UNIVERSITY OF OXFORD

Intelligence as “cognitive skills”



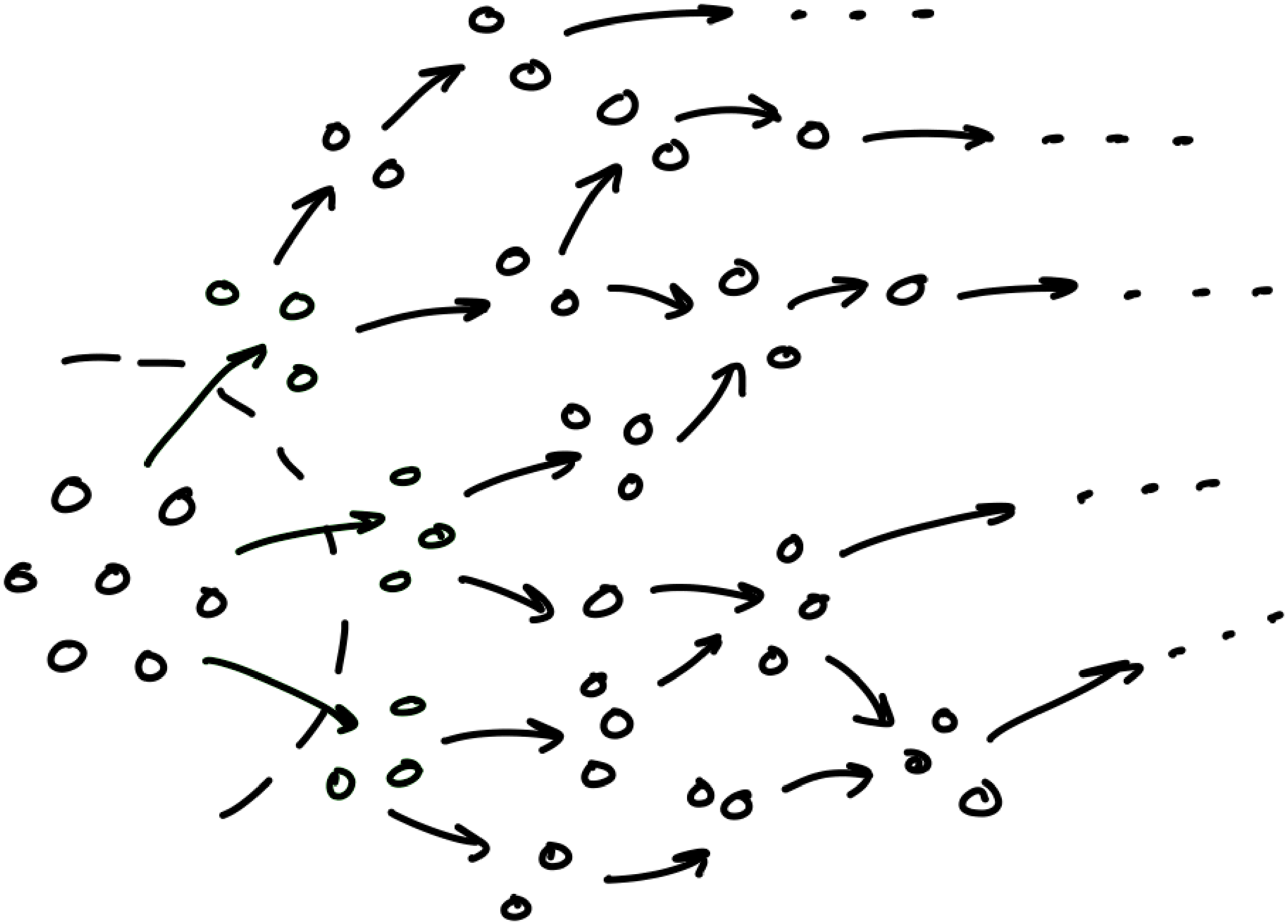
Self-improvement

design of systems with new cognitive skills



Intelligence explosion

fast, repeated self-improvement to super-intelligence



Catastrophic accident pathway:

1. super-powerful inference & planning

2. accidental misuse

3. convergent instrumental goals *(self-improvement, resource acquisition, self-preservation, etc.)*

4. global side-effects *(infrastructure proliferation, threat neutralization)*

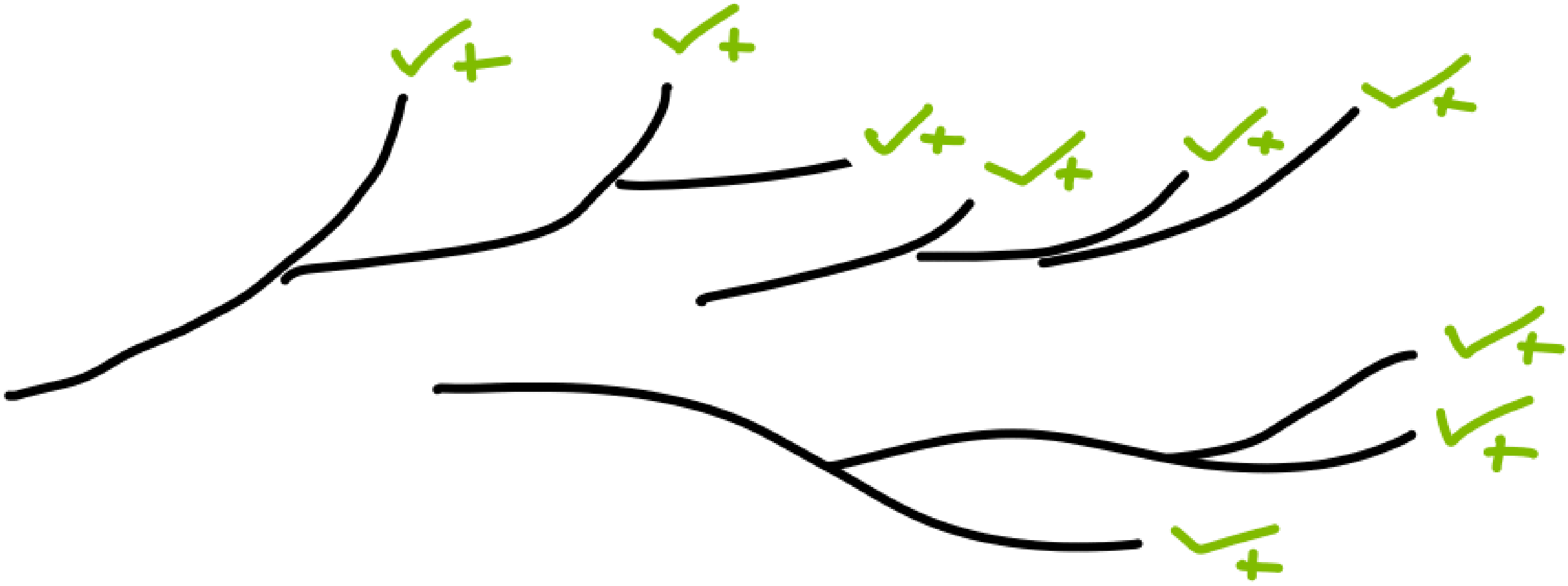
Over many uses, accidental catastrophe via misuse becomes likely

Supposing this, what could we do?

Option A: Future solutions

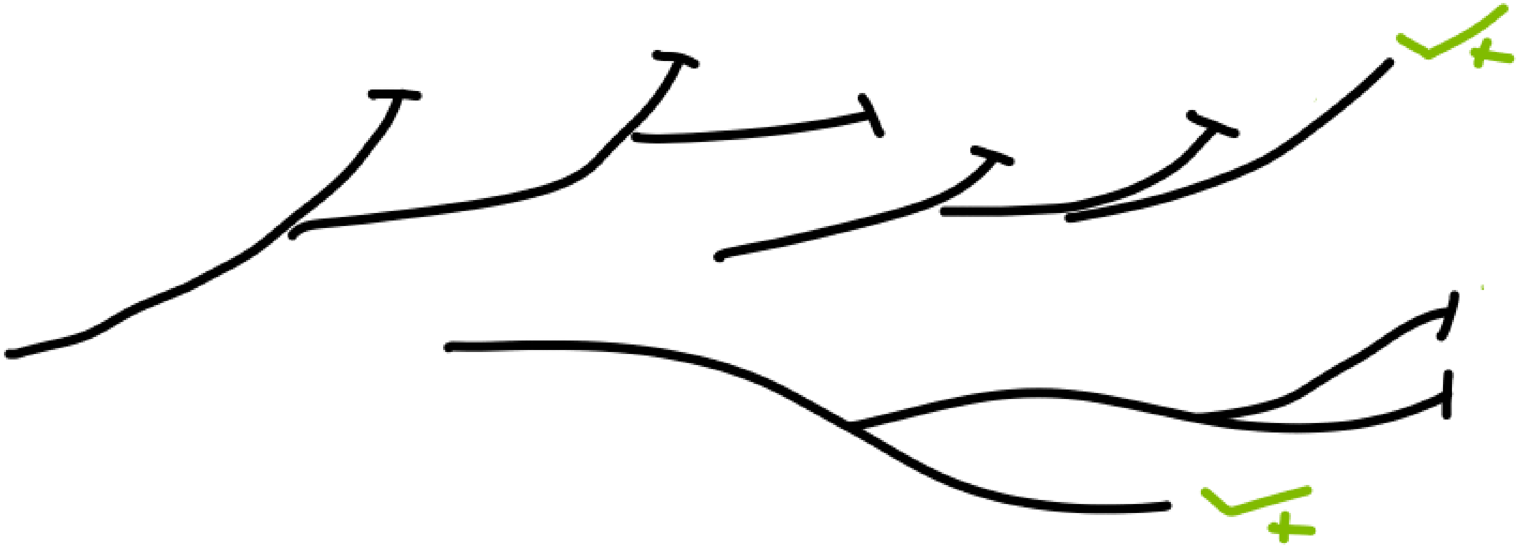
1: Safety engineering

Reduce risk of catastrophic misuse to acceptable levels



2: Regulation of (some kinds of) AI

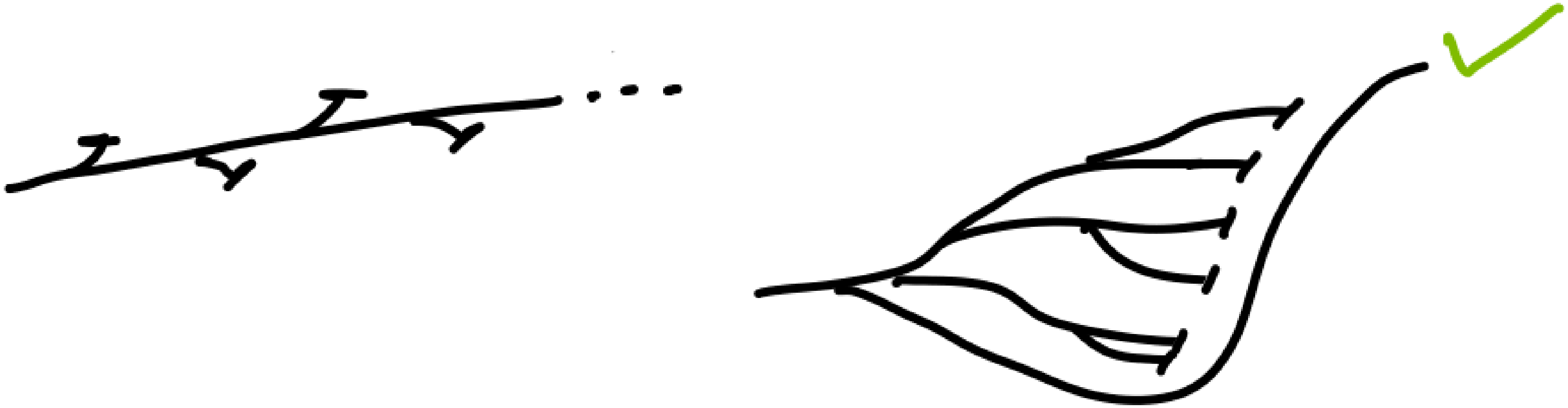
Centralize, control, or otherwise regulate research or use of some kinds of AI



3: Radical solutions

E.g.:

- Extreme regulation (surveillance)
- Controlled, humane-valued explosion



4: ?

Plan out & enact complete solutions:
probably too hard

We lack sufficient information

- …about intelligence explosion
- …about future AI
- …about future society

Option B: Act incrementally to improve future people’s chances of avoiding accidental misuse

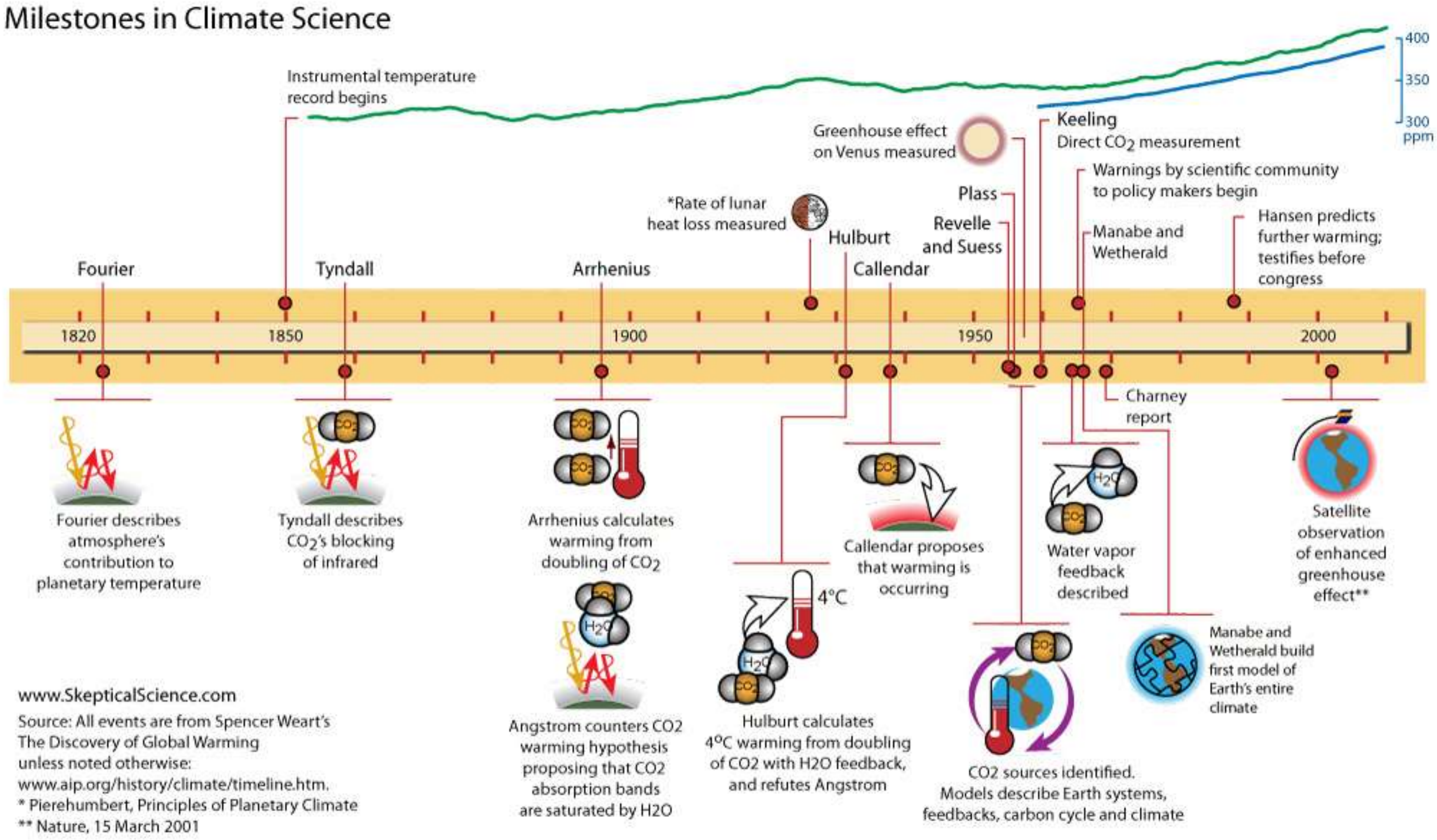
How well-informed are future people about the risks?

How coordinated are future people w.r.t. this issue?

What technical safety knowledge do future people have?

Mostly scientifically addressable! (with a dash of technological risk management policy)

Digression: climate science


Milestones in Climate Science
Instrumental temperature record begins
Greenhouse effect on Venus measured
Keeling
Direct CO2 measurement
400
350
300 ppm
Warnings by scientific community to policy makers begin
*Rate of lunar heat loss measured
Plass
Revelle and Suess
Hulburt
Callendar
Manabe and Wetherald
Hansen predicts further warming; testifies before congress
Fourier
Tyndall
Arrhenius
1820
1850
1900
1950
2000
Charney report
Fourier describes atmosphere's contribution to planetary temperature
Tyndall describes CO2's blocking of infrared
Arrhenius calculates warming from doubling of CO2
Angstrom counters CO2 warming hypothesis proposing that CO2 absorption bands are saturated by H2O
Hulburt calculates 4°C warming from doubling of CO2 with H2O feedback, and refutes Angstrom
4°C
Callendar proposes that warming is occurring
Water vapor feedback described
CO2 sources identified. Models describe Earth systems, feedbacks, carbon cycle and climate
Manabe and Wetherald build first model of Earth's entire climate
Satellite observation of enhanced greenhouse effect**
CO2
H2O

www.SkepticalScience.com

Source: All events are from Spencer Weart's The Discovery of Global Warming unless noted otherwise: www.aip.org/history/climate/timeline.htm.

* Pierehumbert, Principles of Planetary Climate

** Nature, 15 March 2001

Lessons

- “Near-sighted”, i.e. non-solution-proposing, work on important problems can be valuable
- Part-time academic work can be critical, especially in the early life of a field
- Simplistic models have long-term value
- The process may take tens to hundreds of years

Intelligence explosion & safety knowledge

1. intelligence explosion
2. powerful inference & planning
3. convergent instrumental goals
4. global side-effects
5. control

1. intelligence explosion:

- more concrete mechanisms;
- better models: what resources are how important?

“more research needs to be done to better define ‘intelligence explosion,’ and also to better formulate different classes of such accelerating intelligences.” *

2. powerful inference & planning:

- how good is possible with how much resources?
- what resources are bottlenecks?

*Interim Report, AAAI Presidential Panel on Long-Term AI Futures, 2009

3. convergent instrumental goals:

- better models of how these arise;
- could these be mitigated or avoided somehow?

4. global side-effects:

- can CIGs be rendered non-global?
- what pathways to harm would be most promising? how thoroughly can we block them?

5. control:

- how reliably can explosions be predicted, prevented, or contained in early stages?
- how predictable and "stable" could an intelligence explosion be?
- could we encode humane values, means to learn humane values, or "domesticity" values?
- what kinds of explosion-resistant systems could be built?

"additional research… on methods for understanding and verifying the range of behaviors of complex computational systems to minimize unexpected outcomes"*

*Interim Report, AAAI Presidential Panel on Long-Term AI Futures, 2009

Summary

Act incrementally to improve future people's chances of avoiding accidental misuse…

…by improving scientific knowledge of intelligence explosion, powerful inference & planning, global side-effects, and control…

…so that future people will be well-informed about risks, able & willing to coordinate, and will have the technical knowledge necessary.