The Future of Robot Ethics

Joanna J. Bryson
Artificial Models of Natural Intelligence
University of Bath, United Kingdom

@j2bryson

From the Abstract

- What is the current reality of AI?
- Are the sciences of consciousness and ethics far enough along that we can predict the consequences of AI?
- What scenarios should we worry about, and which should we seek to accelerate?

- What is the current reality of AI?
    - It's here now, changing the world.
- Are the sciences of consciousness and ethics far enough along that we can predict the consequences of AI?
    - Yes.
- What scenarios should we worry about, and which should we seek to accelerate?
    - Give me forty minutes...

Definitions

- Definition: How I am going to use a word for the next two hours.
- Aim: to communicate clearly about natural and artificial intelligence,
    - not to describe or debate current "ordinary language" usages.

Intelligence

Culture

Ethics

Consciousness

Moral

Altruistic

Self-Aware

Suffer

Cooperative

Intentional

Soul

Autonomous

Human

Agent

Robot

Definitions

- Definition: How I am going to use a word for the next two hours.
- Will normally choose the simplest meaning that has computational efficacy – can describe a change in information and/or behaviour.
- Aim: to untangle concepts previously confused by historic correlation.

(dis)claimer

- Most of what I’m saying today derives from the scientific literature, can be found in textbooks.
- A few slides towards the end are my own research, will flag them.
- Tweet, email or Google me (or “AI Ethics”) for references.

Cooperative

Culture

Outline

Intelligence

Consciousness

Suffer

Moral

Altruistic

Self-Aware

- Computing & AI Concepts
- Biological & Sociological Concepts
- Psychological & Philosophical Concepts
- Futures

Ethics

Soul

Human

Agent

Autonomous

Robot

Intentional

- Intelligence: the ability to generate appropriate behaviour in response to an unpredictable environment.
- What makes this hard: Computing and Tractability.

DODDER SMELLS ITS HOST;
LOCATES BY VOLATILES FROM HOST

Plants can wind & unwind (reversing decisions) in pursuit of support, light, prey. (Anthony Trewavas, Edinburgh)

Laws of Computation

- Computing: systematically altering the form of information.
- Takes time, energy, and space (memory).
  - Not abstract, not math, not eternal – computing happens IRL.

Why is it hard to be smart?

- Pretend you bought a robot, and it came with 100 things it knew how to do without being told.
- For example, eat, sleep, turn right, turn left, step forward, step backward, pick things up, drop them...
- Now pick a goal for your robot.
- For example, flying to Tokyo.

Sanyo robot watchdog

The hardness of smartness (2)

- Suppose you can't be bothered to tell your robot exactly how to get to Tokyo, so you have it guess.
- If getting to Tokyo takes one step, the robot may have to try 100 different things.
- If it takes two steps, the robot may have to try each thing after each thing:

$$100^2 = 10{,}000$$

The hardness of smartness (3)

- If the robot doesn't know how many steps it takes to go to Tokyo, it might get caught in an infinite loop.
- For example, it might eat, sleep, work, eat, sleep, work, eat, sleep, work... and never buy a passport.
- When computer scientists say "hard" they mean "pretty much intractable."

Sony SDR-4Xs. Pictures from BBC

Tractability

- There are more possible short chess games than atoms in the universe.
- Biology has a **lot** more options than chess.
- Tractability is the fundamental challenge of intelligence.

Intelligence: search for the next action.

Rational cannot mean perfect, optimal, correct

Rationality must be bounded; requires trading off costs, including time and space required for computation.

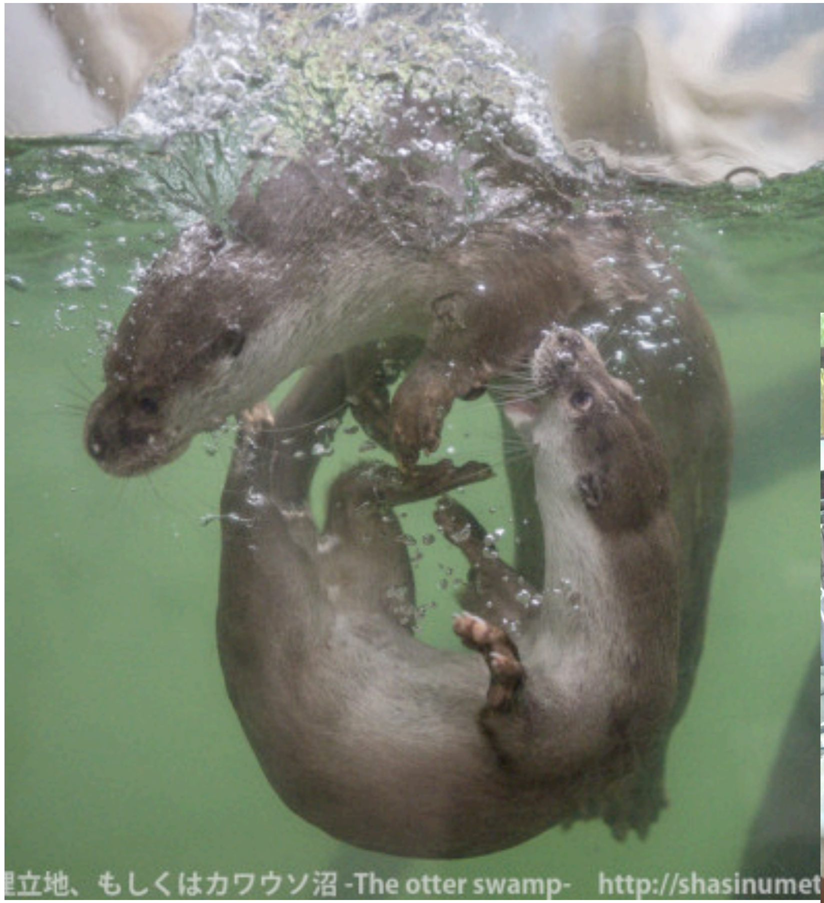
Strategies for Speeding Thought (Search)

- Concurrency
  - multiple searches at the same time,
  - only effective if solutions can be communicated.
- Pruning
  - limit search to likely space of solutions

Search as a Tree that Needs Pruning

- Start with the present as your "root".
- Think of every act you can take: branching.
- From each branch: the next acts as twigs.
- Prune (don't consider) branches you know you won't use (e.g. where you're sure to lose).

Bodies Prune

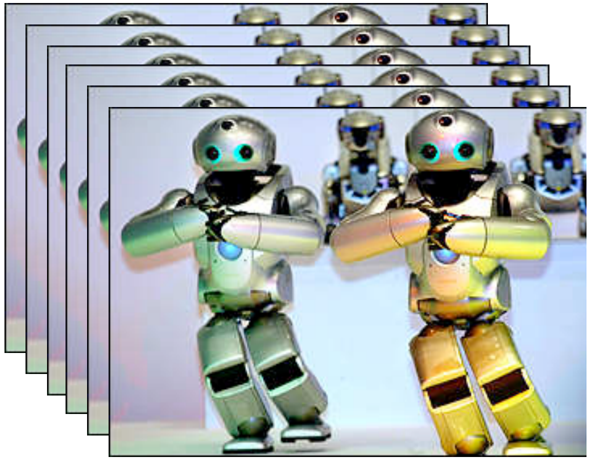


Strategies for Speeding Search

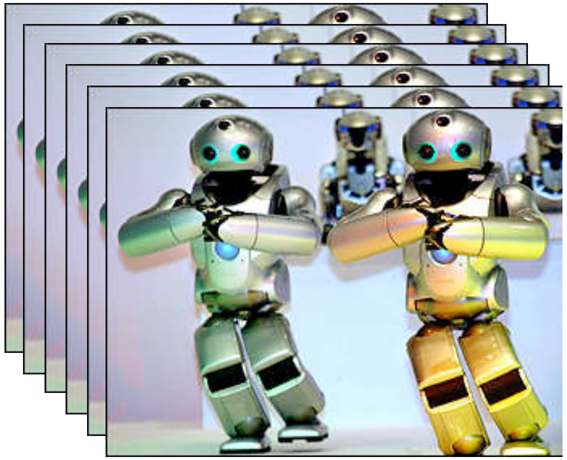
- Concurrency
  - multiple searches at the same time,
  - only effective if solutions can be communicated.
- Pruning
  - limit search to likely space of solutions

Culture as Concurrency

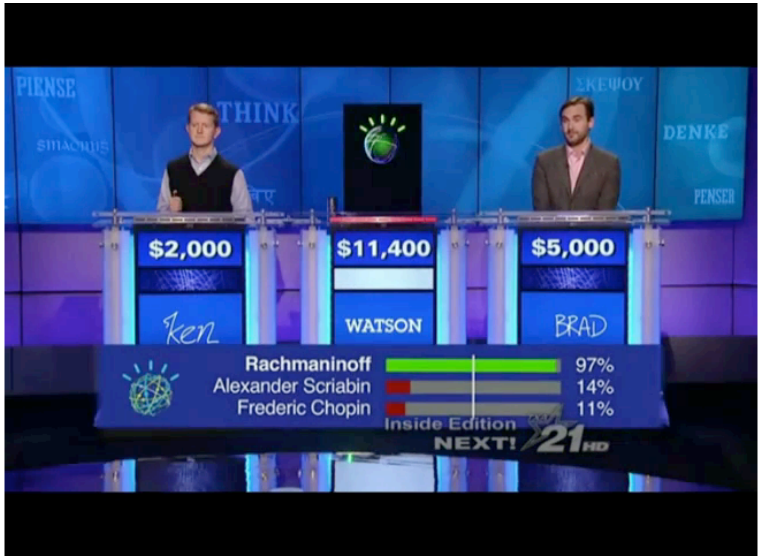
- If every agent has a 1% chance of discovering a skill in its lifetime (e.g. making yogurt), & there are 2000 agents, then at any instant some agents probably have that skill.



- Biological evolution: concurrent search for a good way to convert energy into replication, “communication” by selection.



- Culture: concurrent search for behaviour communicated **within** lifetime.



We are succeeding at AI because we've learned to exploit the discoveries of evolution and culture.

Plus: cloud computing.

Cooperative

Culture

Outline

Intelligence

Consciousness

Suffer

Moral

Altruistic

Self-Aware

- Computing & AI Concepts
- Biological & Sociological Concepts
- Psychological & Philosophical Concepts
- Futures

Ethics

Soul

Human

Agent

Autonomous

Robot

Intentional

Computation, Tractable, Rational, Pruning, Concurrency

- Robot: an intelligent machine that's actions impact the real world in real time, and are based on sensing the real world in real time.

So that's plants and people, right?

No.

Because we build robots and determine their goals, their pruning. Our complete authorship is fundamentally different from our relationship to humans or other evolved systems.

photos: Georgio Metta (top) & Emmanuel Tanguy

Social Behaviour

- Cooperating: behaving to the advantage of others.
- Mutualism: when everyone benefits from cooperation equally.
- Altruism: paying a net cost to benefit others.

Can we find altruistic behaviour in…

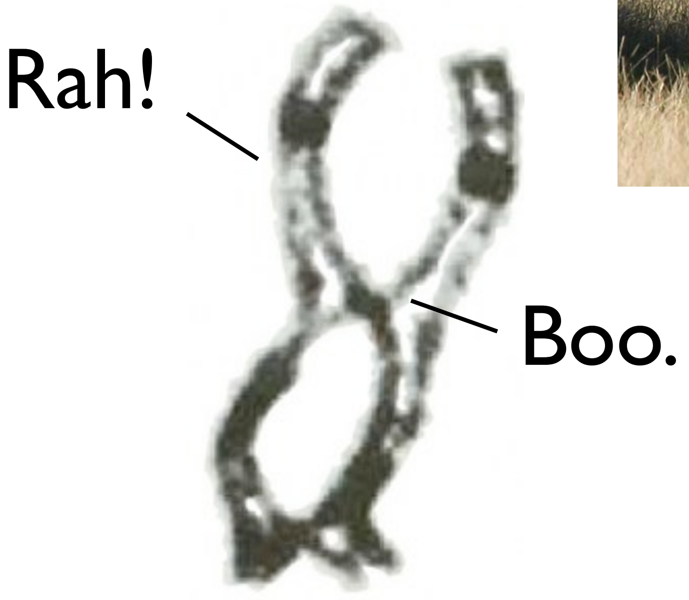
- Robots?
  - Yes, we can program them to make any cost/benefit assessment we choose.
- Nature?
  - Yes, it’s ubiquitous.

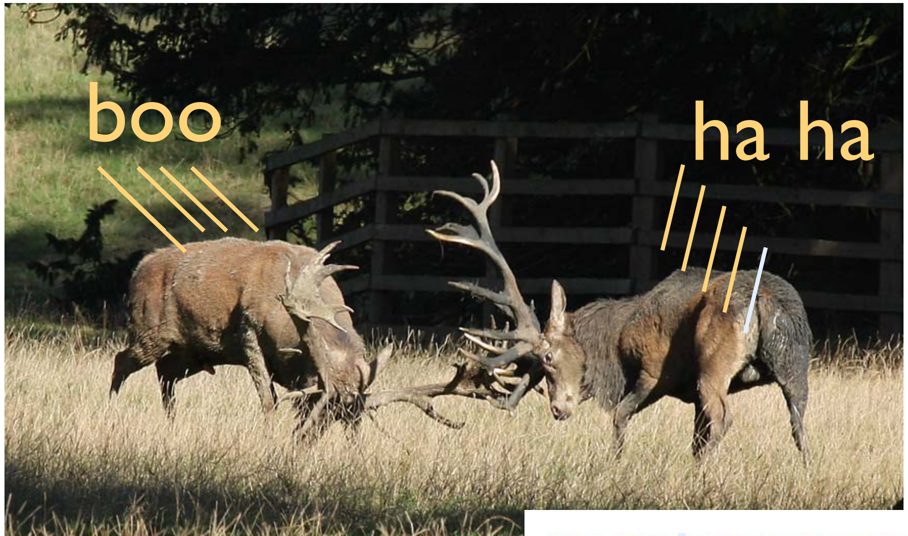
Selfish Genes $\not\Rightarrow$ Selfish Individuals

- Traits advantageous to the community but costly to the individual were (for some time) considered inaccessible to evolution. This is false.
- Explanation: inclusive fitness & kin / group selection
  - What is transmitted is the replicator.
  - The unit of selection is the vehicle (or interactor.)
  - In the current ecology, most vehicles are composed of many, many replicators.

Multiple Levels of Interaction ⇒Cooperation

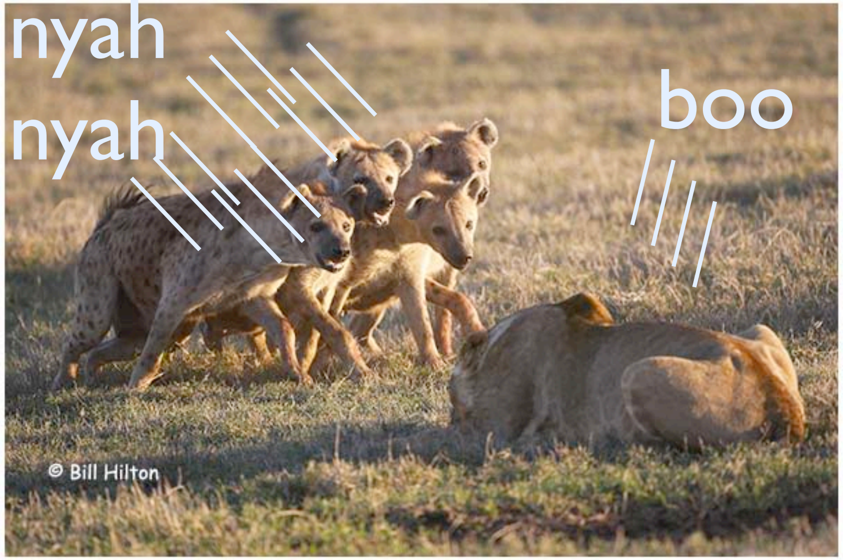
Replicator (Gene)





Organism

Group



So why aren’t we all nice all the time?

- Cooperation is only rational when it increases the probability of replicators persisting.
- Sometimes there’s only enough food for one monkey, or one family, or one village…
- Ethics: systems of behaviour that optimise social living – reduce costs from conflict, possibly favour the group.

Something from my research.

Cooperative

Culture

Outline

Intelligence

Consciousness

Suffer

Moral

Altruistic

Self-Aware

- Computing & AI Concepts
- Biological & Sociological Concepts
- Psychological & Philosophical Concepts
- Futures

Ethics

Soul

Human

Agent

Autonomous

Robot

Intentional

Computation, Tractable, Rational, Pruning, Concurrency

Quick Reminder

- What is the current reality of AI?
- Are the sciences of consciousness and ethics far enough along that we can predict the consequences of AI?
- What scenarios should we worry about, and which should we seek to accelerate?

Agency

- Agent: something that causes change (e.g. chemical agents.)
- Autonomy: Generating some behaviour due to own internal motivations.
- Moral agent: something society considers responsible for its actions.
- Moral patient: something society has responsibility towards.

Next bit is my research / published opinion.

Moral actions require…

- a behavioural context to afford more than one possible action for the individual,
- at least one available action be considered by a society to be more socially beneficial than the other options, and
- the individual is able to recognise which action is socially sanctioned, and able to act on this information.

Easy to build in robots!!
(includes monkeys & pets!)

Displacement of Responsibility

- at least one available action is considered by a society to be more socially beneficial than the other options, and
- Pros & cons of considering the robots responsible?
  - Alternative: considering them intelligent prosthetics of a responsible human’s will, like owned dogs & horses, children.

Joanna J. Bryson “Patiency Is Not a Virtue: Suggestions for Co-Constructing an Ethical Framework Including Intelligent Artefacts”, *The Machine Question: AI, Ethics and Moral Responsibility*, D. Gunkel, S. Torrance and J. J. Bryson (eds.), AISB, Birmingham, 2012.

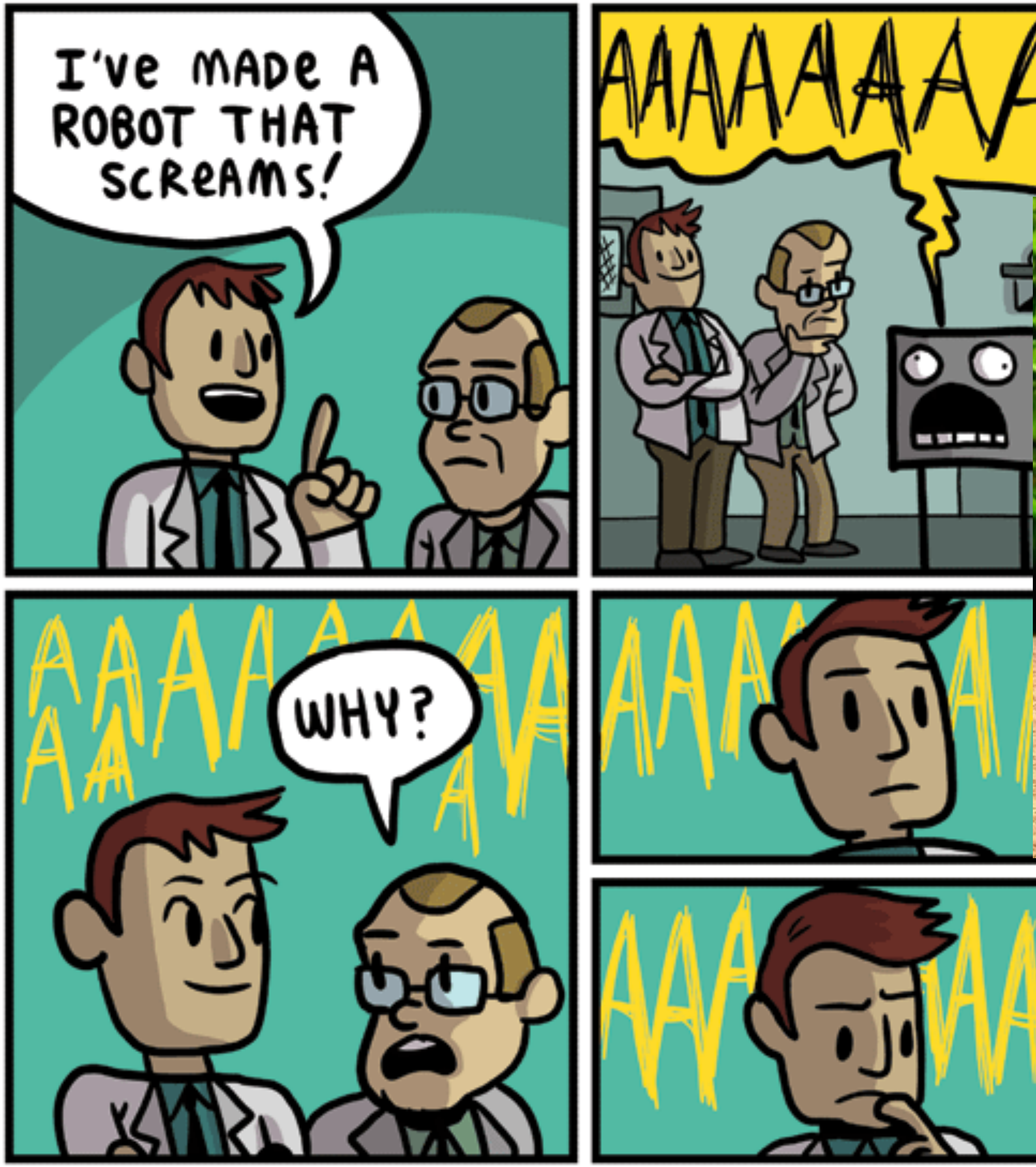
- For Human Society (us):
  - Pros: feel godlike, culture might persist beyond planetary limits, might produce more useful tools.
  - Cons: political & commercial moral hazard, misattribution of blame / resources.
- For AI (them robots):
  - No Pros: (except maybe for the unbuilt).
  - Cons: compete w/ humans for resources, stress of social dominance, fear of death etc.

My Conclusions

We are ethically obliged to make robots we are not ethically obliged to.

Deeming robots to be moral agents unethically neglects our responsibility as authors of their intelligence.

Note: these are normative assertions, not facts.



KC Green, *Gun Show*, #513

What if they're conscious?

What I Think Consciousness Is

- A specific kind of attention. Name for the feeling you have as you evaluate chains of actions with intermediate outcomes. Required for learning new plans.
- Language (just) helps: culture accumulates concepts that are likely to help you focus conscious attention on worthwhile things.
  - "Self" is one of those concepts.

Consciousness for AI

- Only need it if system learns and learning relies on a bottlenecked cognitive resource.
- If you need it, allocating it to tasks you are doing in proportion to how uncertain you are about them is a pretty good guess.
    - Also attend to other novel / unpredicted by your internal model events.

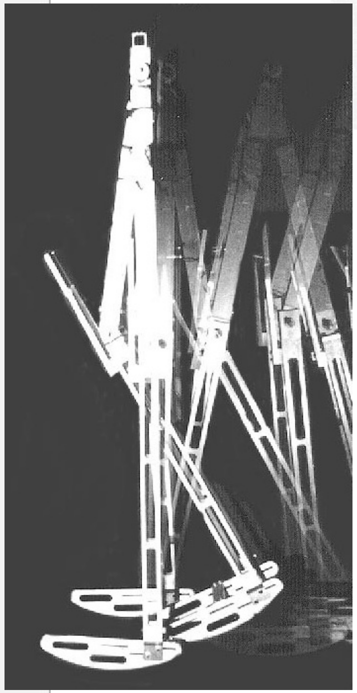
Joanna J. Bryson "A role for consciousness in action selection", *International Journal of Machine Consciousness* **4** (2):471–482, 2012.

Joanna J. Bryson, "Age-Related Inhibition and Learning Effects: Evidence from Transitive Performance", in *Proc. of the 31st Annual Meeting of the Cognitive Science Society (CogSci 2009)* pp. 3040–3045, July 2009.

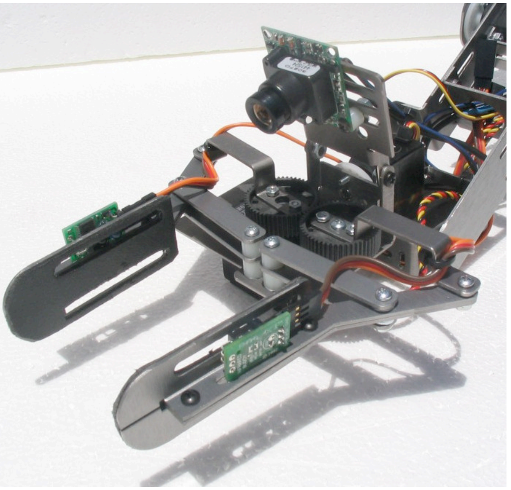
What's Consciousness?

it is nor
hand, nor
foot, nor
arm, nor
face, nor
any other
part
belonging
to a man.

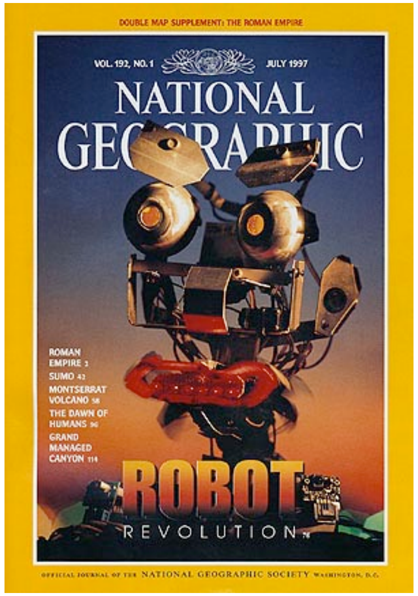
Glenn Matsumura, Wired 2007



Tad McGeer's passive dynamic walker

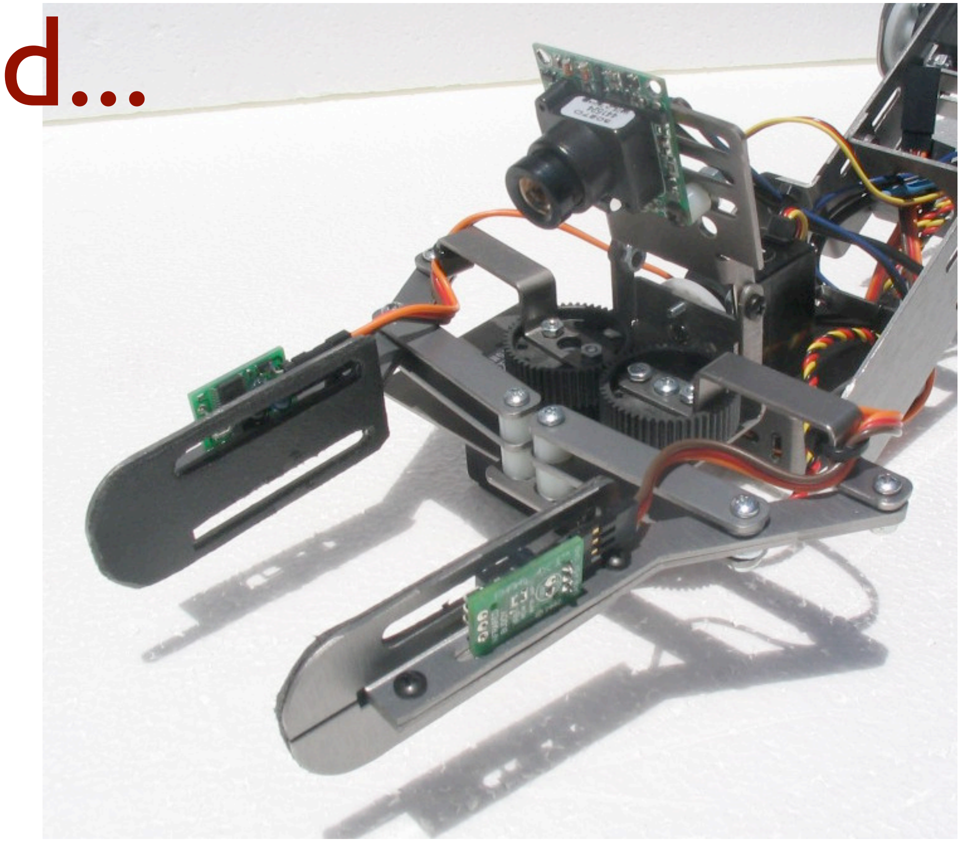


SG5-UT Robotic Arm



Chuck Rosenberg's IT, 1997

If this is a hand...



…then aren't these conscious systems?

Leonardo can now generalize the task goal to a new configuration

Andrea Thomaz, MIT

Charlie Kemp, GA Tech

Correlation ≠ Causation

- In humans, we are only responsible of what we are aware of.
- But awareness isn't magic (though it is the basis of all our intentional communication.)
- Unawareness: a form of pruning.
- Awareness: a subpart of our intelligence that helps us learn complex relations.
- Intentional: actions we were aware of.
- Soul: supposed property making you a moral patient.

Cooperative

Culture

Outline

Intelligence

Consciousness

Suffer

Moral

Altruistic

Self-Aware

- Computing & AI Concepts
- Biological & Sociological Concepts
- Psychological & Philosophical Concepts
- Futures

Ethics

Soul

Human

Agent

Autonomous

Robot

Intentional

Computation, Tractable, Rational, Pruning, Concurrency

AI & Society

is does not imply ought

descriptive ≠ normative

AI ethics relates two types of human artefact: ethical systems & robots.

Normative not descriptive ethics: there is no pre-determined slot for AI to discover.

Question: is there any utility in displacing the responsibility we as authors have onto AI?

Not a question: whether it's possible.

Futures

- What I most want:
  - Sustainability, lack of suffering & conflict.
  - Better regulation of our resource exploitation and distribution.
- What I most fear:
  - Allowing that regulation to reduce individual variation, eccentricity helps learning.

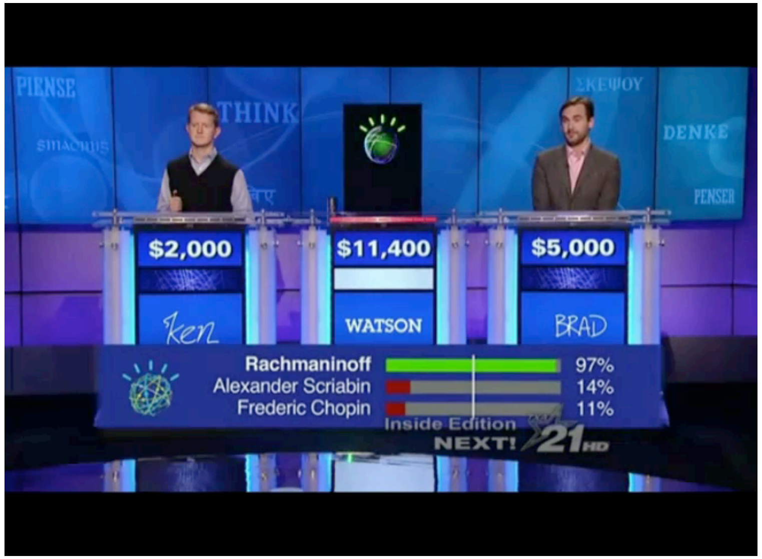
Intelligence (AI, Culture, Governance) increases both.

Recommendations

- Don't make robots or AI unnecessarily humanoid.
  - Increases the chance we allocate resources, responsibility to it inappropriately.
- Treat our personal data like our homes.
  - Only lease information for specific purposes.
  - Anyone can break in, but anyone who does should go to jail.

AI Already Owns Our Advantages

Utopia: Solve hard problems like sustainability; reliably supporting everyone’s efforts to self actualise.



Dystopia: Losing autonomy / ability to freely express; catastrophic disruption of the global ecosystem.

Thanks!

... and other collaborators

My current students:

~~Daniel Taylor~~

~~Bidan Huang~~

Dominic Mitchell

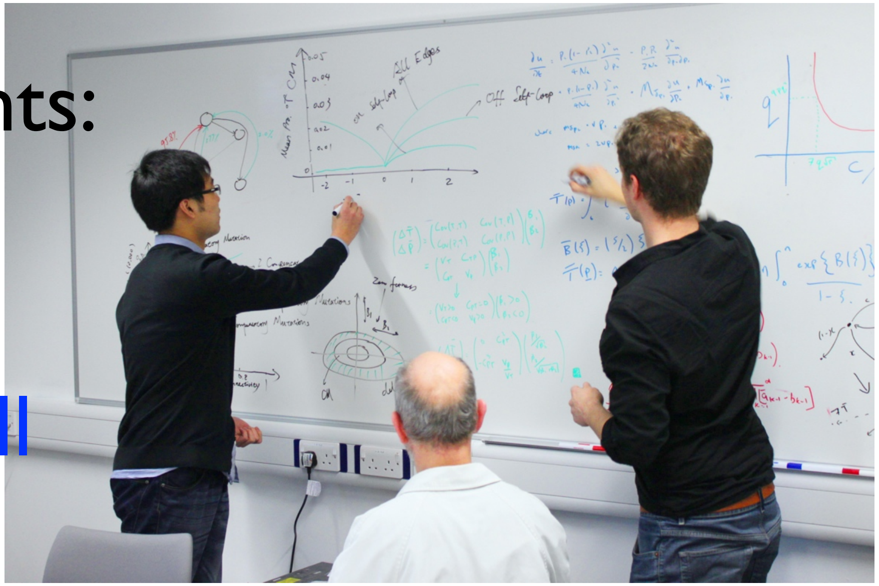
Swen Gaudl

Paul Rauwolf

Jekaterina Novikova

Yifei Wang

Rob Wortham



Will Lowe

Ivana Čače

Mark Wood

Benedikt Herrmann

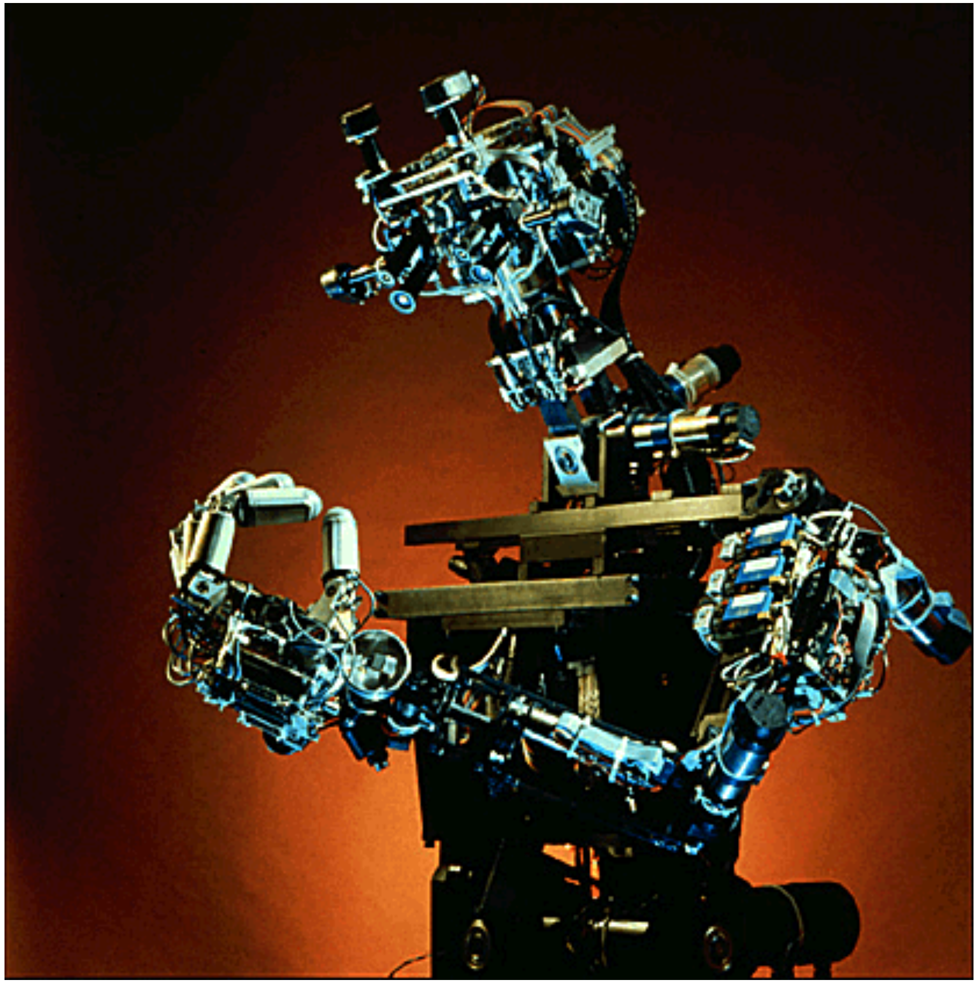
James Mitchell

Simon Powers

Phillip Rohlfshagen

Karolina Sylwester

Manu Tanguy



People want to make AI they owe obligations to, can fall in love with, etc. – “equals” over which we have complete dominion.

Joanna J. Bryson and Philip P. Kime, “Just an Artifact: Why Machines are Perceived as Moral Agents”, *The Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, Barcelona, Spain, pp. 1641–1646, Morgan Kaufmann, 2011.

I still suspect it matters that we 'makers' know how our robots work.

For makers, AI is like a movie.

We can immerse ourselves with characters we know are fiction.

But for others, could it be deception?

photo from: Tony Belpaeme

Humans Are Just Chimps With History

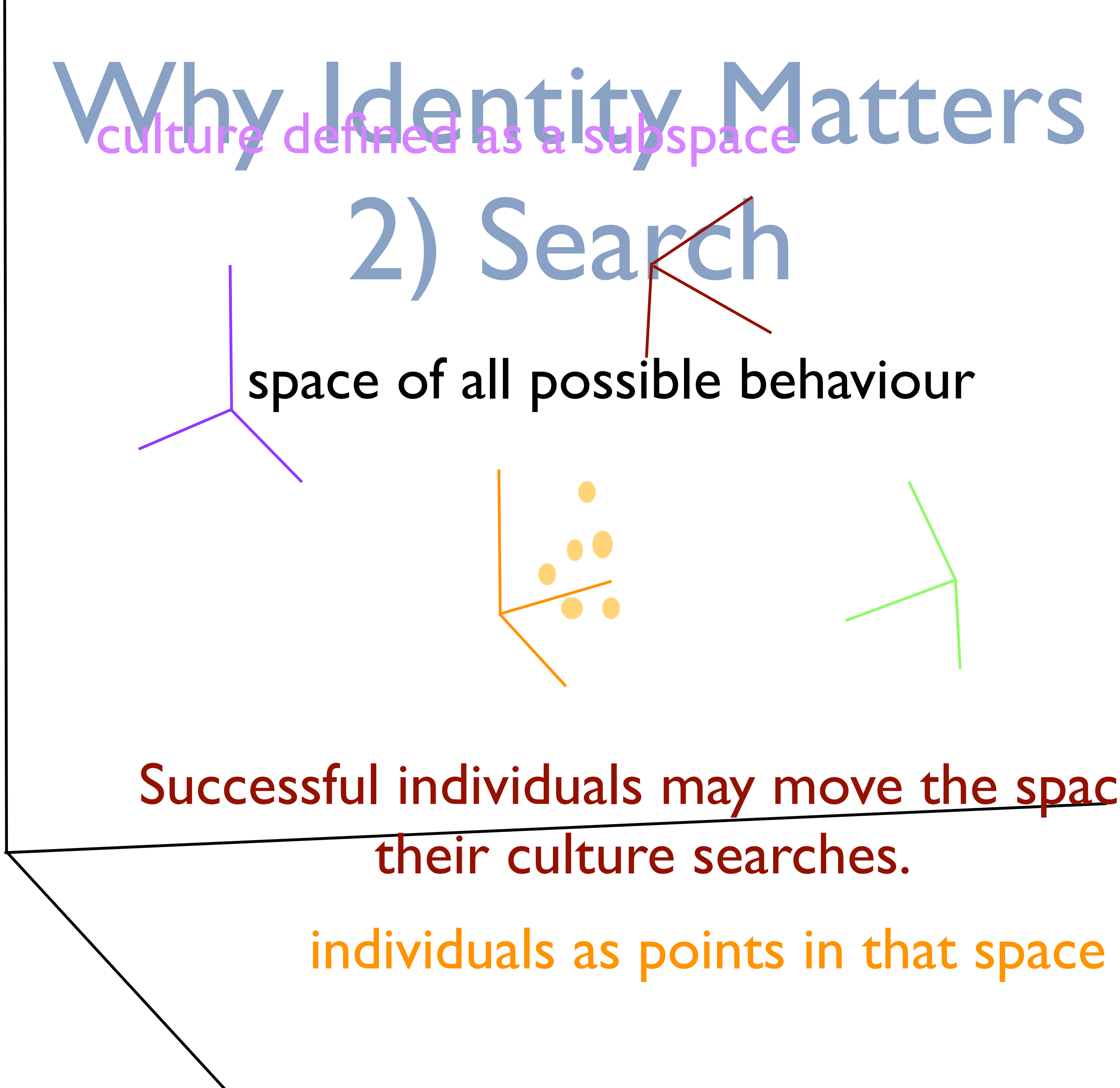
- A brain 3x the size of a chimp’s is not a big deal.

- Humanity only became **really** different in the last 10,000 years, with writing, urbanisation, doctrinal religion (spreading knowledge) & agriculture.

Origins of Ethics

- Ethical systems regulate our sociality.
- Presumably, ethical systems coevolve with our sociality.
- Sociality involves both co-existence of individuals and identity as a group (which co-exists with other groups.)

Why Identity Matters
2) Search



culture defined as a subspace

space of all possible behaviour

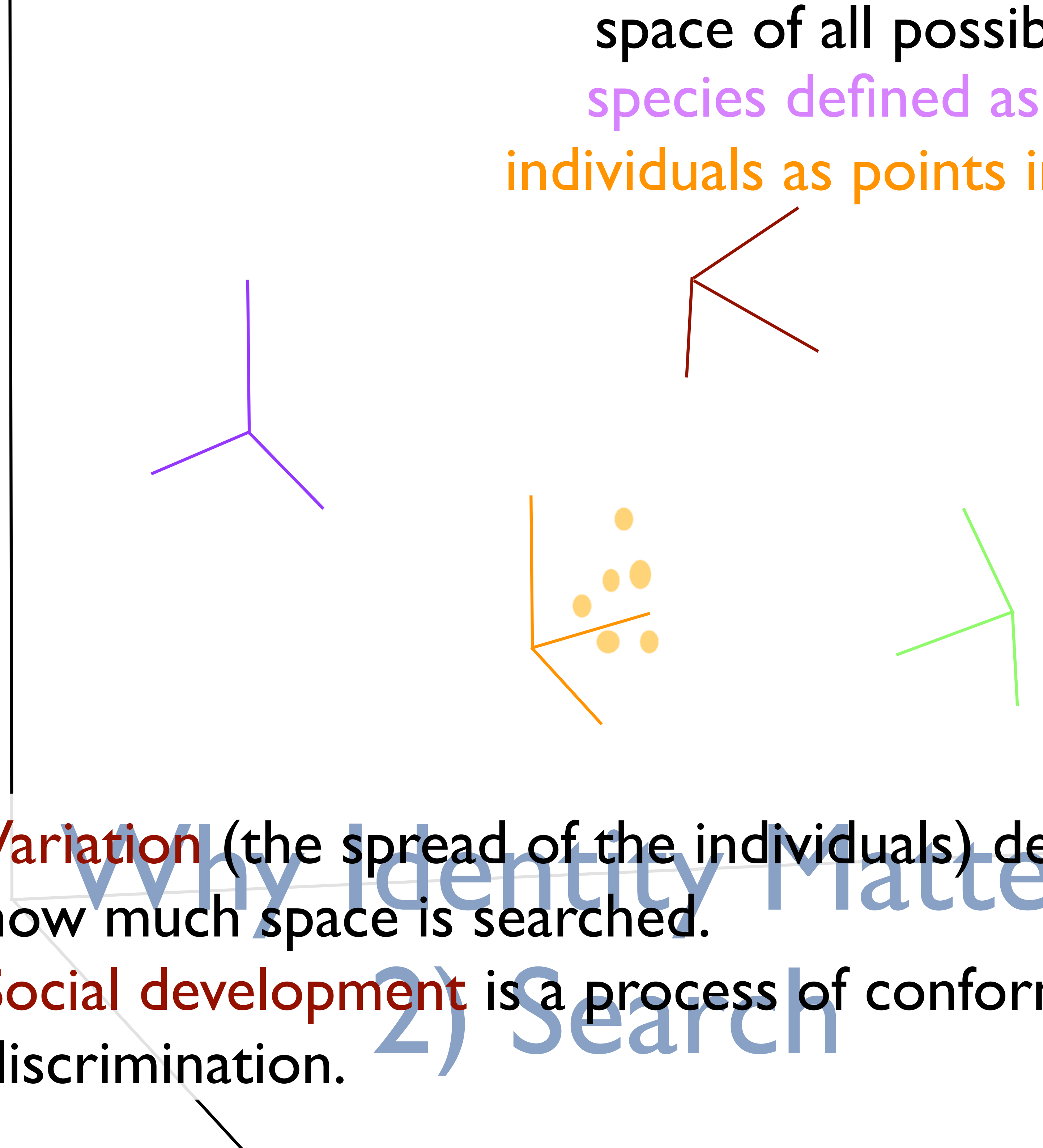
Successful individuals may move the space their culture searches.

individuals as points in that space

space of all possible behaviour

species defined as a subspace

individuals as points in that space



Why Identity Matters
2) Search

- Variation (the spread of the individuals) determines how much space is searched.
- Social development is a process of conformity and discrimination.

Roadmap for Conscious Machines

Arrabales, Ledezma, & Sanchis, 2009

1. (-1) Disembodied
1. (0) Isolated
1. Decontrolled
2. Reactive
3. Adaptive
4. Attentional
5. Executive
6. Emotional
7. Self-conscious
8. Empathic
9. Social

10.Human-like

11.Super Conscious

Roadmap for Conscious Machines

1. (-1) Disembodied

1. (0) Isolated

1. Decontrolled

2. Reactive Sensing to action: *intelligence*

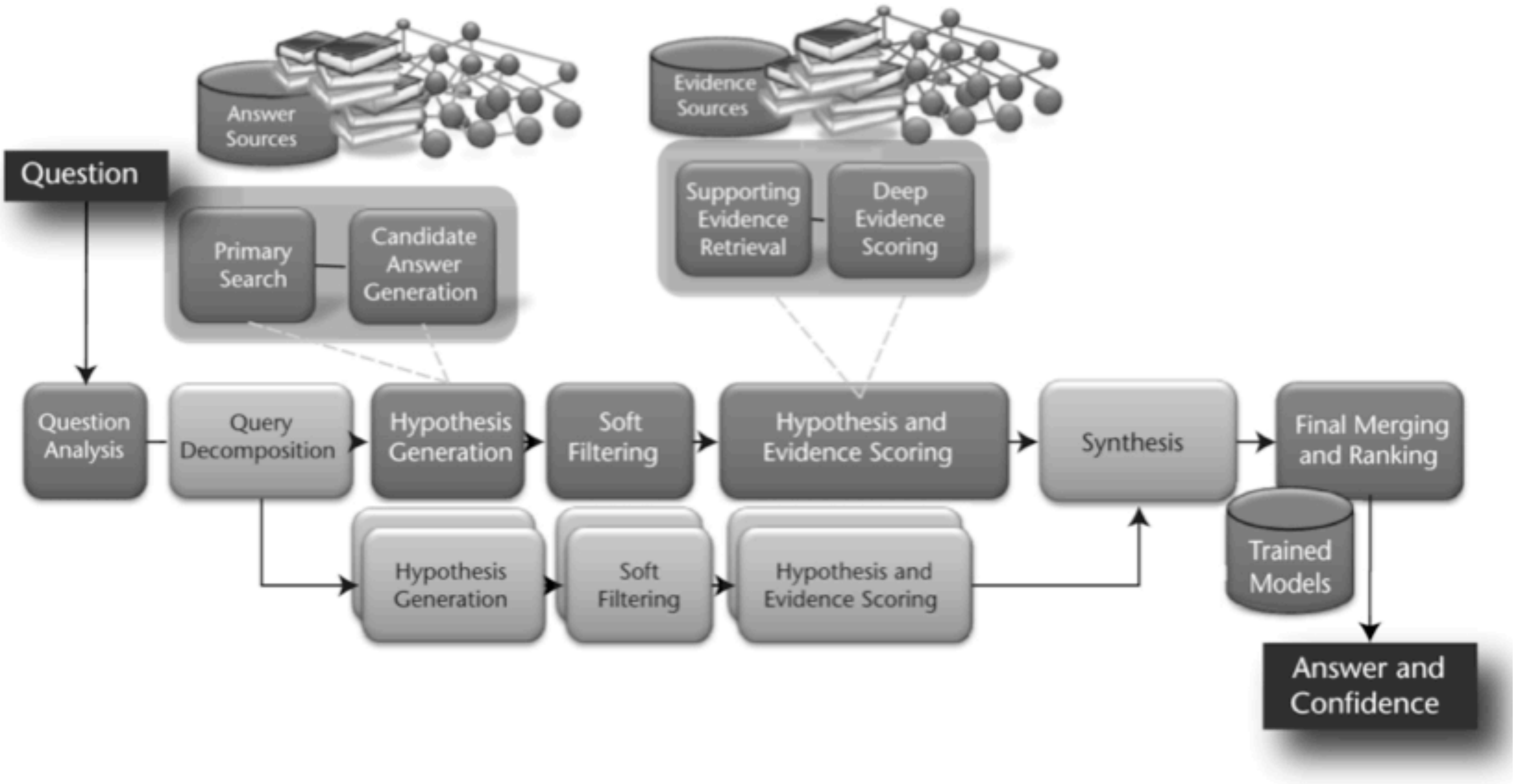
3. Adaptive Learning

4. Attentional Unconsciousness is more conscious!

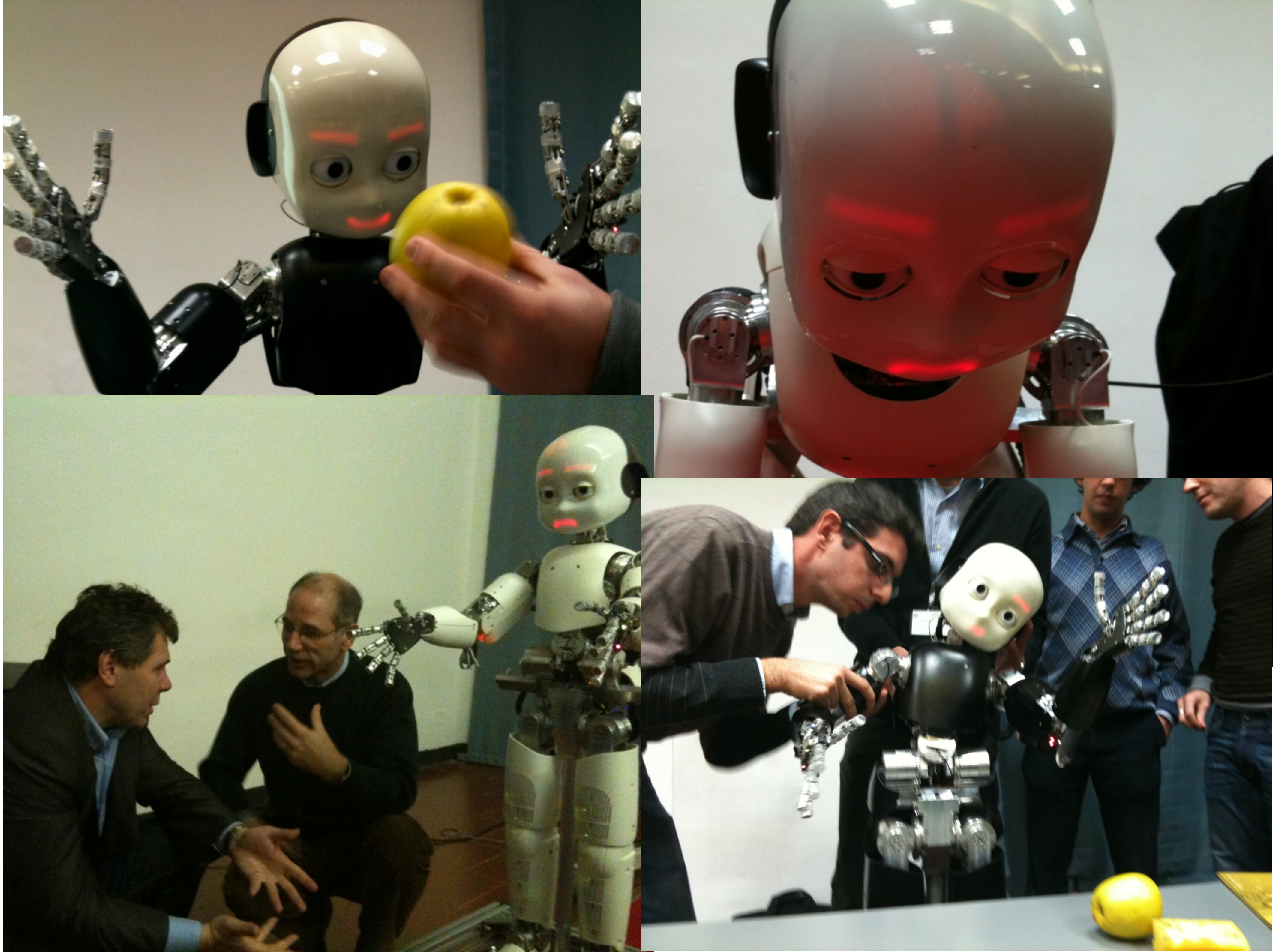
Arrabales *et al.* 2009

Roadmap for Conscious Machines

5. Executive	multiple goals (unconscious 2)
6. Emotional	“human like” (???)
7. Self-conscious	knows about self
8. Empathic	knowledge (k) of others
9. Social	k of other’s k of self
10. Human-like	use Interweb to extend mind
11. Super Conscious	multiple streams!



(Ferrucci *et al.*, AI Magazine 2010)



We Get to Decide...

- Whether robot minds are unique or have autosave and offsite backup.
- Whether robots suffer (permanently degrade their behaviour) when neglected, insulted or otherwise subordinate.
- We as authors are ethically obliged to ensure we have no ethical obligations to AI.

UK EPSRC's Principles of Robotics (2011)

1. Robots are multi-use tools. Robots should not be designed solely or primarily to kill or harm humans, except in the interests of national security.

2. Humans, not robots, are responsible agents. Robots should be designed & operated as far as is practicable to comply with existing laws & fundamental rights & freedoms, including privacy.

3. Robots are products. They should be designed using processes which assure their safety and security.
(of 5...)

UK EPSRC's Principles of Robotics (2011)

4. Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent.

5. The person with legal responsibility for a robot should be attributed. [like automobile titles]

Joanna J. Bryson, Kerstin Dautenhahn and Geoff Pegman, "Man and the machine", letter published online, *The Economist*, 16 June 2012.
Joanna J. Bryson "The Making of the EPSRC Principles of Robotics", *The AISB Quarterly*, (133) Spring 2012.

Big Dog (by Boston Dynamics)

Jeopardy vs Watson

Videos via Dale Lane, IBM