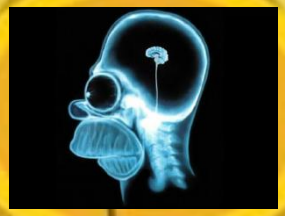
Cognition in Artificial Agents

- History of GofAI (Good old fashioned AI)
  - The ups and downs of a field!
- Problems
- Embodied Cognition – whatever that is?
- Some thoughts and the way towards some solutions



Early attempts at formal reasoning:

- Aristotle (384 BC – 322 BC “Syllogism”)
- Euclid (300 BC, “Elements”)
- al-Khwārizmī (c. 780– c. 850, Algebra.He gave his name to "algorithm")
- William of Ockham (c. 1287 – 1347, Occam’s Razor)
- Ramon Llull (1232–1315, logical “machines “ described as mechanical entities)

Philosophers of „Reason“

- Leibniz, Hobbes (reasoning is "nothing more than reckoning“) and Descartes led to the idea of Physical Symbol Systems

> "A physical symbol system has the necessary and sufficient means for general intelligent action." — Allen Newell and Herbert A. Simon

Mathematicians

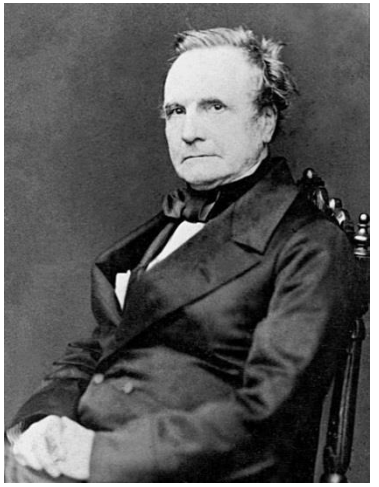
- Gödel's incompleteness proof, Turing's machine and Church's Lambda calculus.

> There are intrinsic limits to mathematical logic, but within these limits, any form of mathematical reasoning can be mechanized.

Syllogism:
Major premise: All humans are mortal.
Minor premise: All Greeks are humans.
Conclusion: All Greeks are mortal.

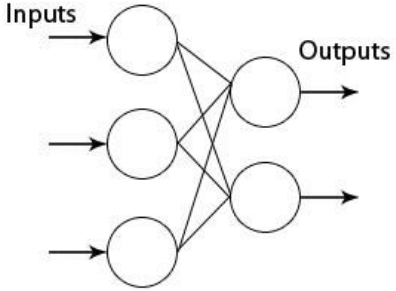
Intelligence and Computers (historical)

- Calculating machines: Charles Babbage (1791 – 1871, programmable computer, Analytical Engine, never built)
- Programs: Ada Lovelace (1815 – 1852) wrote a set of notes that completely detail a method for calculating Bernoulli numbers with the Engine.



Computers versus Networks

- Real Computers: Second WW Code-breakers, Z3, ENIAC, Colossus



Neural Networks

- Walter Pitts and Warren McCulloch (idealized artificial neurons, first artificial neural network.
- Marvin Minsky (first neural net machine, SNARC)
- Connectionism: Parallel distributed processing (e.g. Perceptron, Rosenblatt, 1958)

Two very different Schools!

Conventional Computers: Symbolic calculations ("Symbol Manipulation"), Symbols and Sentences store the information (human readable)

Connectionist Approaches: Implicit calculations, Synapses store the information (not human readable).

AI, its birth and the early years

- Birth of AI: The Dartmouth Conference of 1956 (Minsky, McCarthy, Shannon and others)

  Core assumption: AI captures every aspect of learning or any other feature of intelligence which an be so precisely described that a machine can be made to simulate it"

- **But** up to the early seventies, the capabilities of AI programs were limited. Even the most impressive could only handle trivial problems .

The Problems:

- Limited computer power: For example Moravec (1976) stated human edge- and motion detection capabilities requires 10^9 operations/second (1000 MIPS). Really we find that practical computer vision applications require 10,000 to 1,000,000 MIPS. By comparison, in 1976 the fastest supercomputer in 1976, Cray-1 did 80 to 130 MIPS.
- Intractability and the combinatorial explosion. Many, even trivial problems are NP-complete (cannot be solved in polynomial, but only in exponential time). Travelling salesman problem (NP-hard!): Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and returns to the origin city?

  The most direct solution would be to try all permutations and see which one is cheapest using brute force search [complexity O(n!)], so this solution becomes impractical even for only 20 cities (better approaches exist, still…….20++……sniff, still small numbers!!)
- Commonsense knowledge and reasoning. Many important artificial intelligence applications like vision or natural language require enormous amounts of information about the world (not even now this is easily available). Related to this is: Moravec's paradox: Proving theorems and solving geometry problems is easy for computers, but a supposedly simple task like recognizing a face or crossing a room is extremely difficult.
- Inherent Logics Problems, (frame problem, qualification problem, ramification problem)

Only around 2000++, finally some true successes:

- Deep Blue (05/1997) beats Garry Kasparov in Chess
- A Stanford robot (2005) won the DARPA Grand Challenge (desert trail drive, 131m).
- A CMU robot won the DARPA Urban Challenge (55m city drive).
- Watson (IBM system) wins Jeopardy! (02/2011) against the two best champions.
- Many commercial successes (Google)

How did this come about – the sad truth!:

- Intelligence as plain number crunching and rigorous (boring) rule adherence using multiple modules!

  (Deep Blue=10^7x faster than Ferranti Mark 1, first chess computer, 1951).

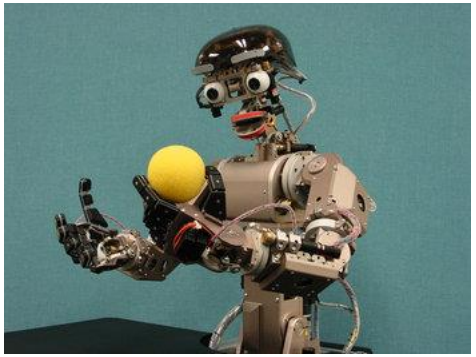
Still there are no truly intelligent, cognitive, and flexible artificial agents (robots) so far.

Where is HAL in 2013 ?

Arthur C. Clarke and Stanley Kubrick (1968) created "HAL 9000", an autonomous computer (believed to exist by the year 2001)

Now move one letter forward for each letter of "H","A","L"

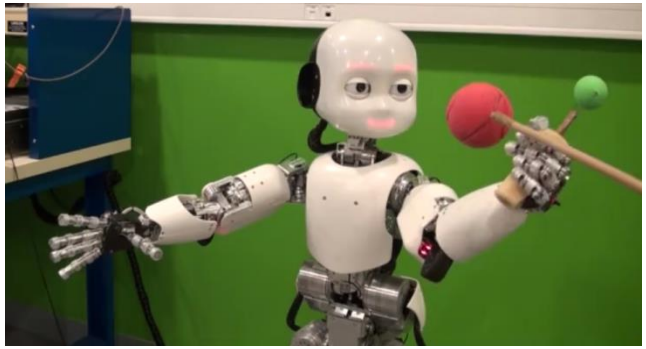
New AI: “Embodied Cognition”



- To show real intelligence, a machine needs to have a body — it needs to perceive, move, survive and deal with the world. Sensorimotor skills are essential to higher level skills like commonsense reasoning.

- One should build intelligence "from the bottom up”. Outside-in, exploration based knowledge acquisition, leads to Developmental Robotics.

- “Elephants Don't Play Chess” (Rodney Brooks, 1990): Symbols are not necessary since "the world is its own best model. It is always exactly up to date. It always has every detail there is to be known. The trick is to sense it appropriately and often enough.



Foundation: “Law of cause and effect”, Thorndike, 1911

Problems and Questions

Hardware

Substrate

Cartesian (knowledge w.o. substrate) versus Embodied (knowledge needs a body)

Data Formats

Computation and storage:

Symbolic (Computer programs) versus Sub-symbolic (Connectionism)

Software

Knowledge acquisition:

Outside-In (exploration based) versus Inside-Out (generative, experience based)

The Weird Stuff

Philosophy of Mind

Hypotheses: Strong AI versus Weak AI

Strong AI Hypothesis: A computer which behaves as intelligently as a person must also necessarily have a mind and consciousness (whatever that is…..Monism, Dualism, etc.).

Problems and Questions

Hardware

Substrate

Cartesian (knowledge w.o. substrate) versus Embodied (knowledge needs a body)

Data Formats

Computation and storage:

Symbolic (Computer programs) versus Sub-symbolic (Connectionism)

Software

Knowledge acquisition:

Outside-In (exploration based) versus Inside-Out (generative, experience based)

The Weird Stuff

Philosophy of Mind

Hypotheses: Strong AI versus Weak AI

Strong AI Hypothesis: A computer which behaves as intelligently as a person must also necessarily have a mind and consciousness (whatever that is…..Monism, Dualism, etc.).

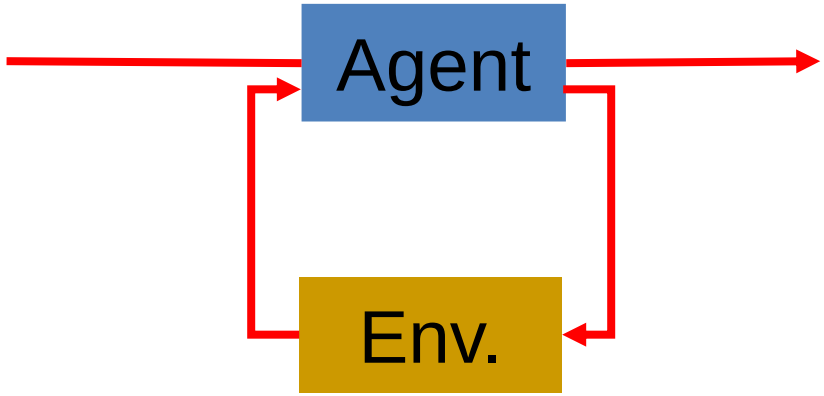
Hardware

Substrate

Cartesian (knowledge w.o. substrate) versus Embodied (knowledge needs a body)

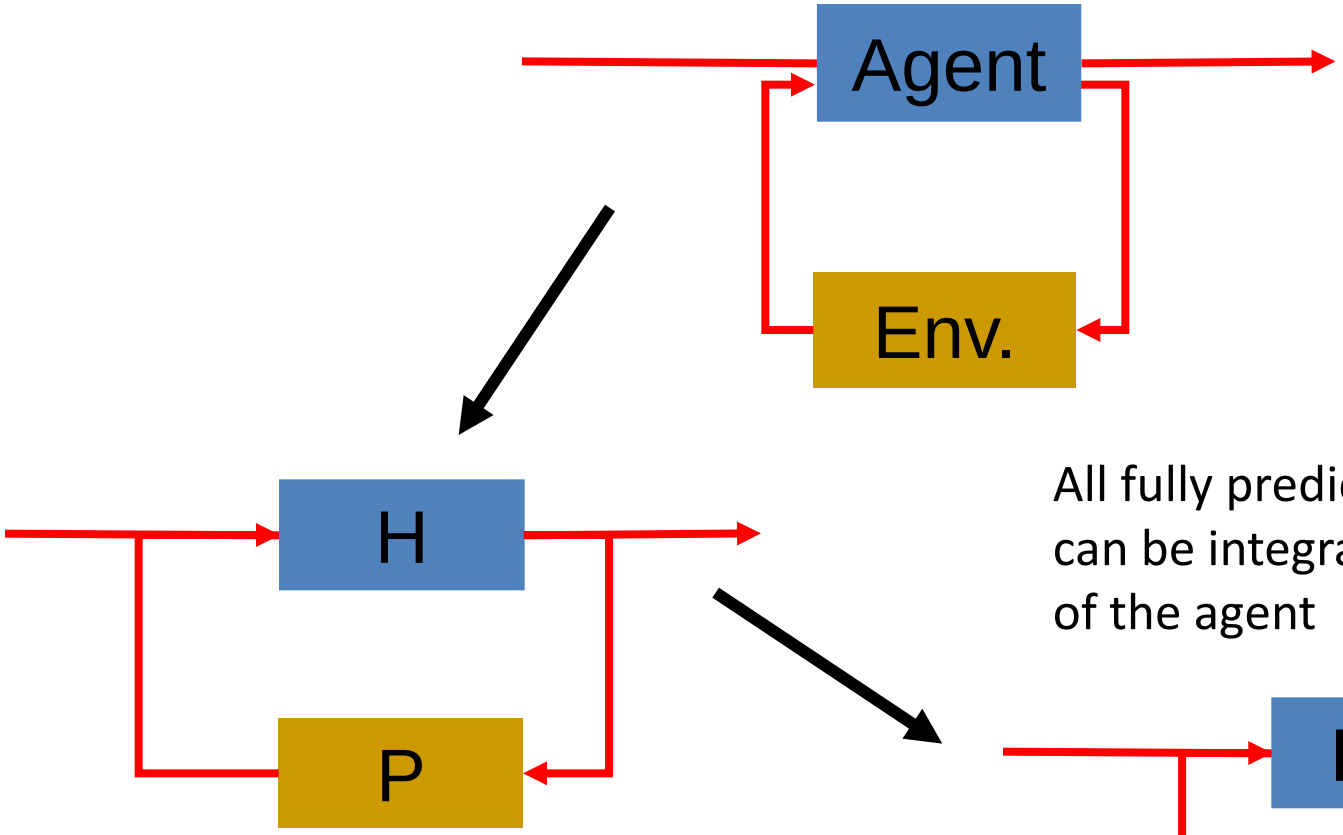
Why would “a body” help an agent to be/become intelligent?

What does it mean to have a body?
What is a body?

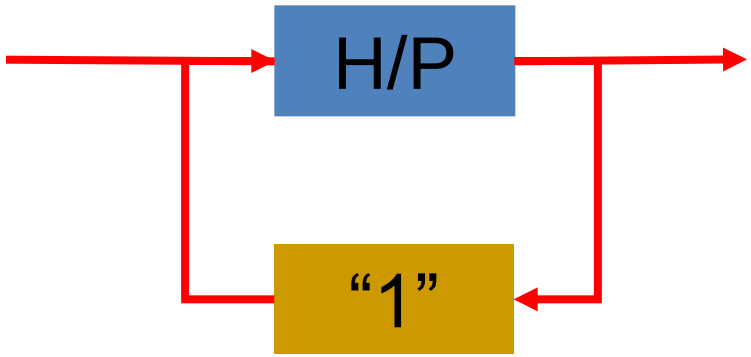


Can only be defined “against” the environment !

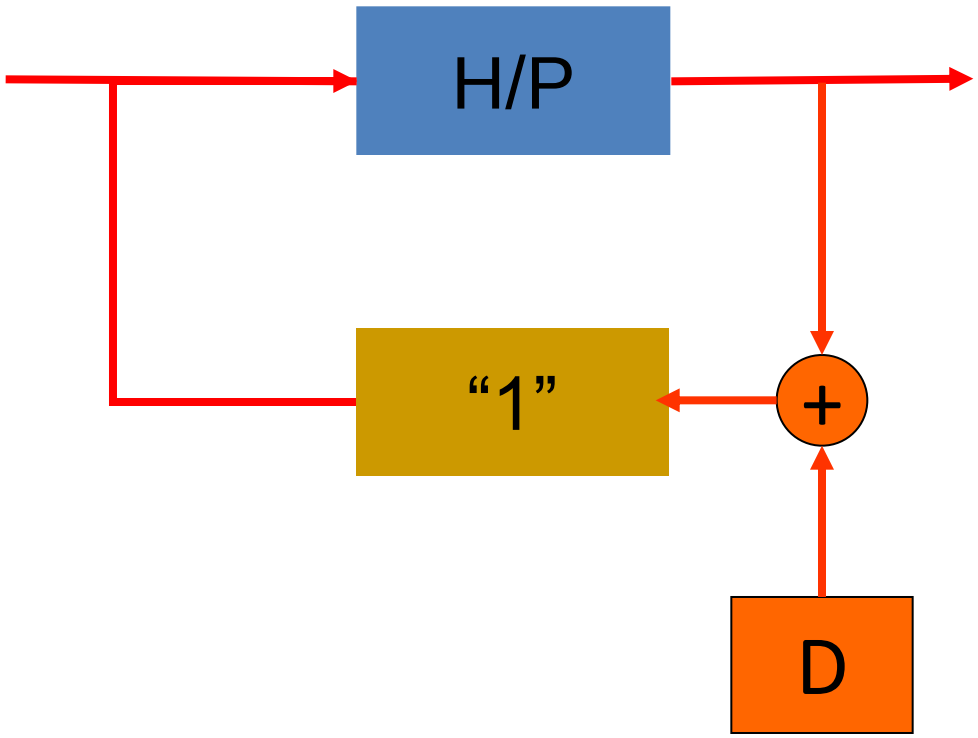
Agents as (linear) Systems



All fully predictable events can be integrated into the body of the agent



On Transfer Functions



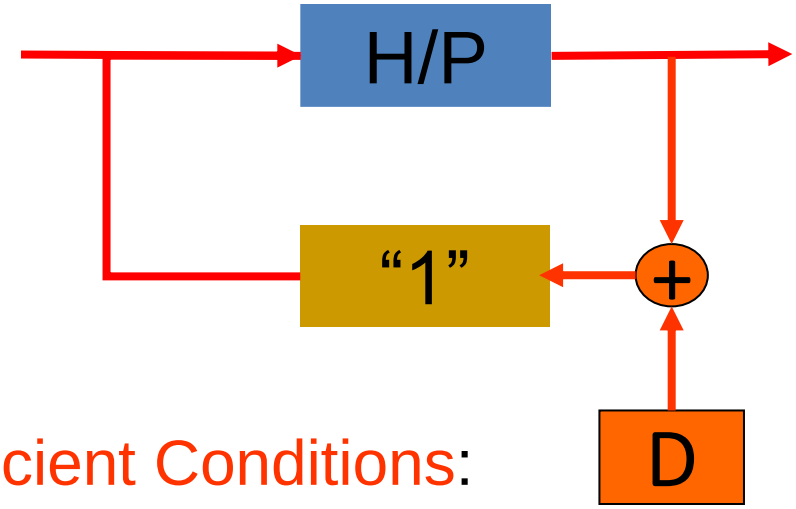
Necessary condition For being part of your body: Every entity who's effects are fully predictable could be part of your body!

Unpredictable disturbances always belong to (come from) the world

Some random examples:
Predictable pain can be to some degree ignored, unpredictable pain not.
Well fitting prostheses can be ignored (bodily integrated).
A race-car pilot becomes "one" with his machine.

Porr, B. and Wörgötter, F. (2005) Inside Embodiment — What means Embodiment for Radical Constructivists? Kybernetes, 34, 105-117.

On Transfer Functions



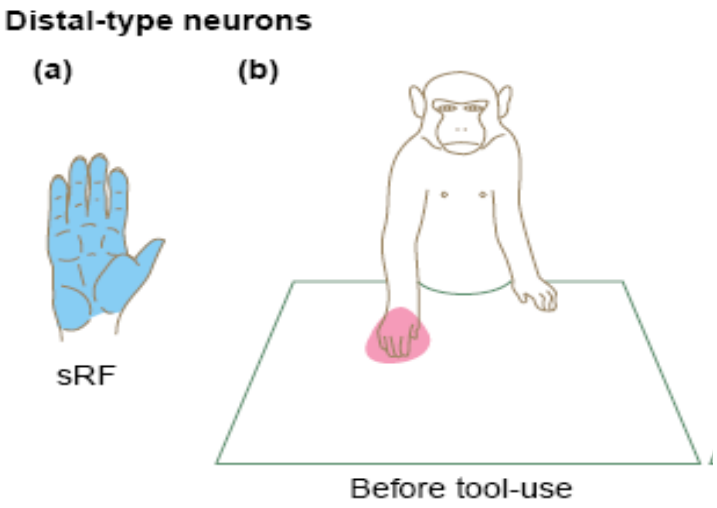
Everything which is fully predictable could be part of your body (Necessary condition)

Sufficient Conditions:

1) To be part of your body the entity, from which a predictable event arises, should be proximal and causally linked to your currently existing body.

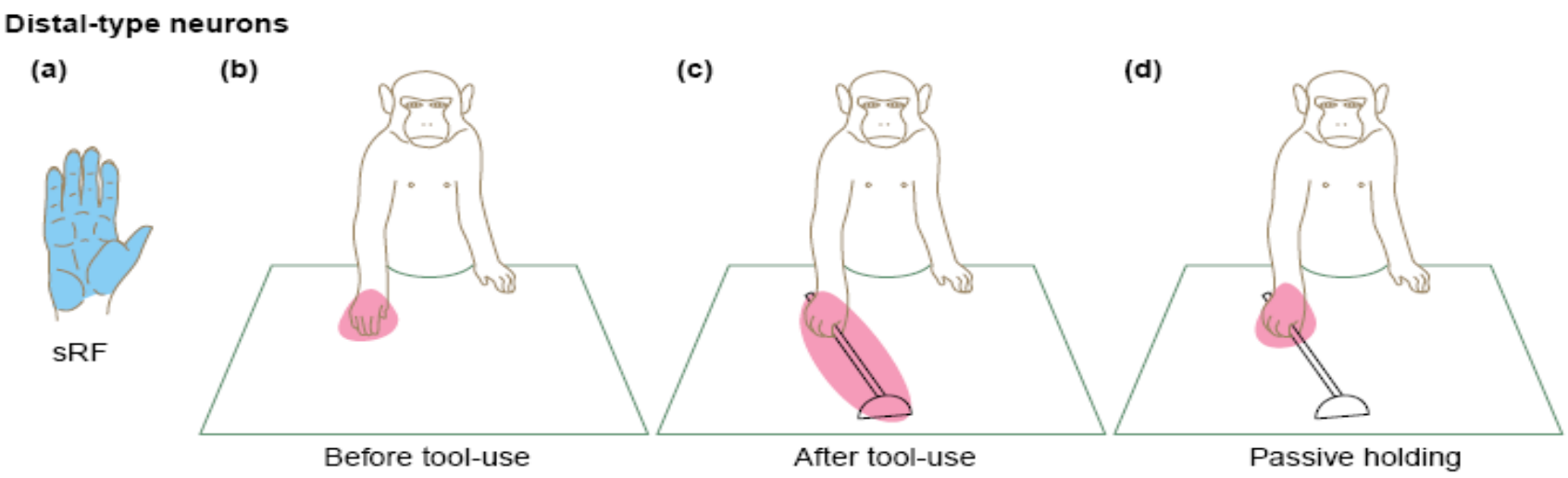
2) To be part of your body any (newly integrated) entity should be part of your body "for a longer time" (Bodies are continuous over some time).

Some examples:
The sun's motion is fully predictable but the sun certainly cannot be integrated into your body.
A robot's hand is linked to a robot's arm.
Two computers are linked by a wireless connection.



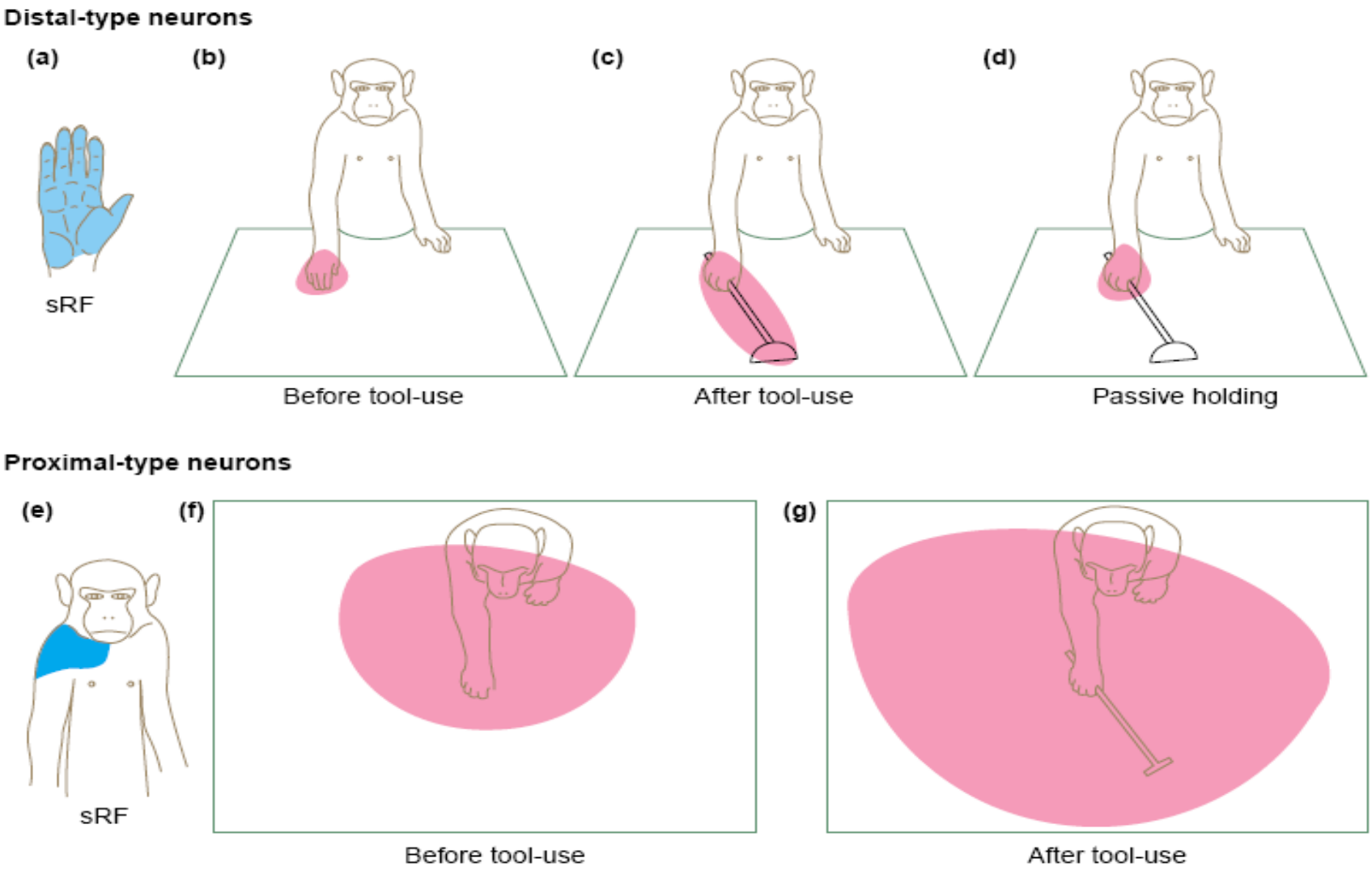
The idea that humans (and monkeys) indeed perform temporary bodily integration is supported by experimental results that over time cortical receptive fields are extended representing the tip of a stick, which a monkey had to use to obtain food for an prolonged period of time.

Obayashi, S., Tanaka, M. and Iriki, A. (2000). Subjective image of invisible hand coded by monkey intraparietal neurons. NeuroReport 11, 3499-3505.



The idea that humans (and monkeys) indeed perform temporary bodily integration is supported by experimental results that over time cortical receptive fields are extended representing the tip of a stick, which a monkey had to use to obtain food for an prolonged period of time.

Obayashi, S., Tanaka, M. and Iriki, A. (2000). Subjective image of invisible hand coded by monkey intraparietal neurons. NeuroReport 11, 3499-3505.

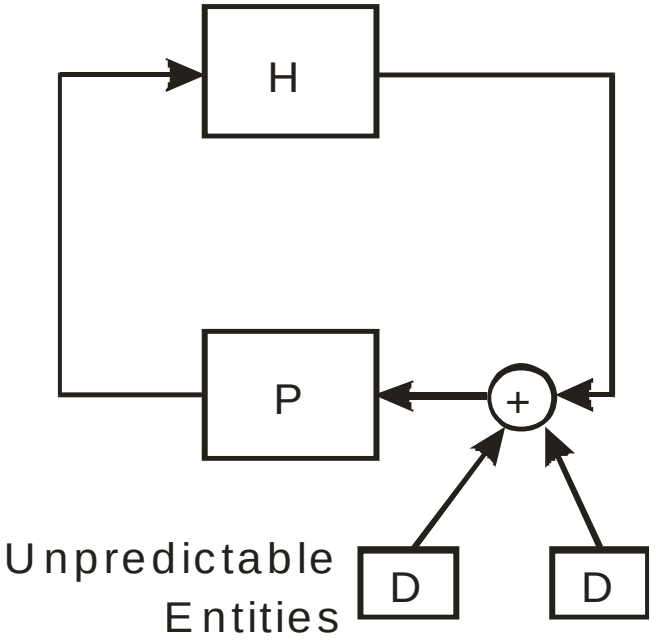


The idea that humans (and monkeys) indeed perform temporary bodily integration is supported by experimental results that over time cortical receptive fields are extended representing the tip of a stick, which a monkey had to use to obtain food for an prolonged period of time.

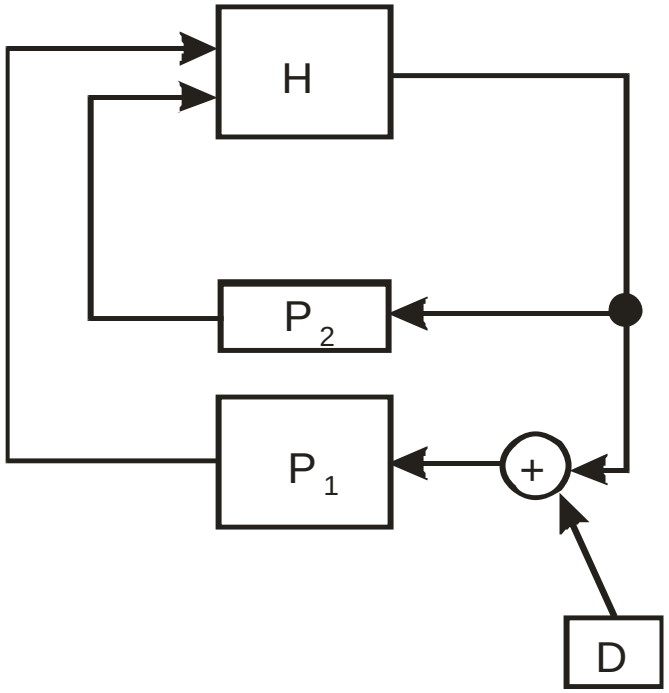
Obayashi, S., Tanaka, M. and Iriki, A. (2000). Subjective image of invisible hand coded by monkey intraparietal neurons. NeuroReport 11, 3499-3505.

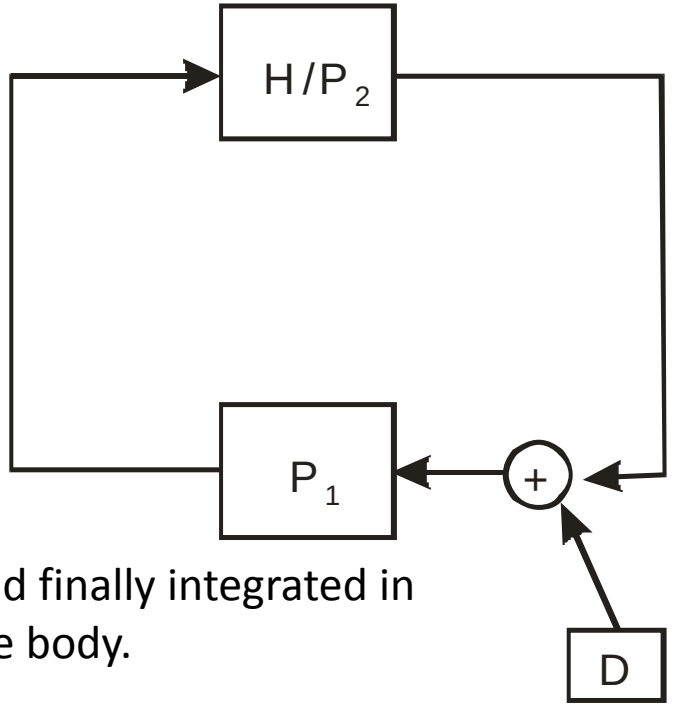
What has happened from a systems theoretical viewpoint?

Bodily integration as a process of integrating a predictable transfer function into the agent's transfer function!



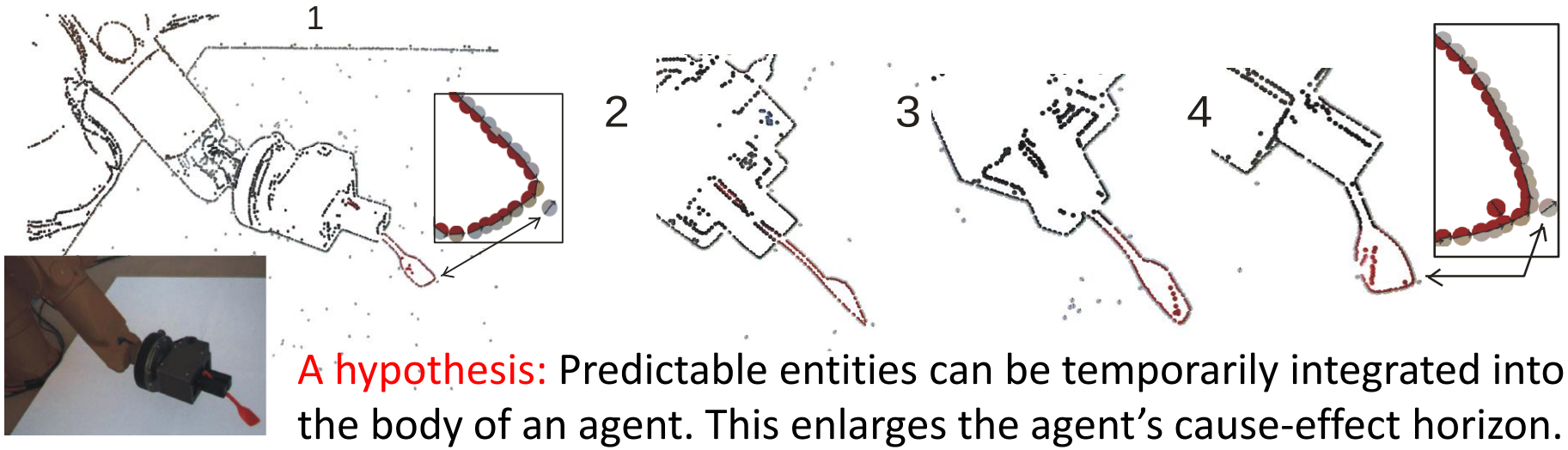
The predictable D can be first regarded as a transfer function in the world





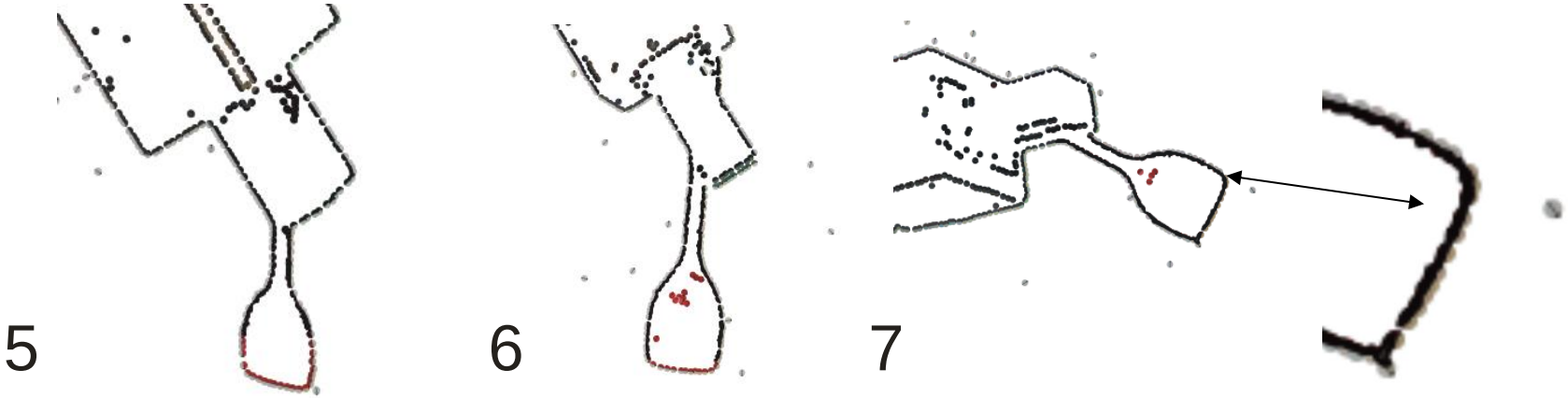
And finally integrated in the body.

Hypothesis: Predictability leads to "body-extension"



A hypothesis: Predictable entities can be temporarily integrated into the body of an agent. This enlarges the agent's cause-effect horizon.

This may well be a strong route to intelligence!



What looks like a simple "re-colouring" really is a difficult computer vision based process of using the RBM principle to "make the spoon part of the robot"

Wörgötter, F., Agostini, A., Krüger, N., Shylo, N. and Porr,B. (2009). Cognitive Agents – A Procedural Perspective relying on "Predictability" of Object-Action Complexes (OACs). RAS, 57(4):420-432

Note: This analysis suggest that the body does not have to be "physical" (material).

Pure internet agents, computer viruses, etc. could fulfill the necessary and sufficient conditions for having a body.

Also: There are quite many embodied systems that are not intelligent. (e.g. Bacteria Professors after they got tenure….

Ok! But what is it that might make them intelligent?

We will argue that it is the degree and complexity of interaction with the world that an agent can entertain!

What happens during an interaction?

For this we need to understand representations, on which the interactions can take place, first.

Problems and Questions

Hardware

Substrate

Cartesian (knowledge w.o. substrate) versus Embodied (knowledge needs a body)

Data Formats

Computation and storage:

Symbolic (Computer programs) versus Sub-symbolic (Connectionism)

Software

Knowledge acquisition:

Outside-In (exploration based) versus Inside-Out (generative, experience based)

The Weird Stuff

Philosophy of Mind

Hypotheses: Strong AI versus Weak AI

Strong AI Hypothesis: A computer which behaves as intelligently as a person must also necessarily have a mind and consciousness (whatever that is…..Monism, Dualism, etc.).

Data Formats

Computation and storage:

Symbolic (Computer programs) versus Sub-symbolic (Connectionism)

"Representations" to achieve intelligence: A problem!

Explicit: By symbols
Implicit: By spike-trains and synapses

Implicit: This works. Just look into a mirror!
But its very hard to copy this!

Explicit: Still, many researchers think that symbolic approaches should be more powerful to arrive at intelligence. So where is the problem?

Symbol Grounding Problem: Symbols are made by us. But really they need to be made by the agent itself.

Thus, next, we will discuss Symbolic Representations

Symbol Grounding Problem: Symbols are made by us. But really they need to be made by the agent itself.

For example: What is an Object?

A container

Traditional feature based representations (edges, color, 3D features, etc.)

What do these items have in common?


Objects suggest Actions!
Affordance Principle, Gibson
(here: „filling“ and/or „drinking“)

But.....

Thus, objects do not exist in their own right!

It is the

specific set of attributes

(required for a certain Handlungsplan, OAC)

that defines the object.

This traditional view is incomplete !

The importance of required attributes is continuous and will vary according to context and need.

Getting water from a desert well

Preferred!

Available

And even this on a rope (in case of need)
























Thus, really ALL these items are „Fillable“!

“Sucking the Sock”

Arguably, there is no set of low or intermediate-level features/attributes whatsoever that captures the „being fillable“ of these vastly different things.

So how can an agent arrive at this?

What makes a thing „fillable“ ?

What makes this thing a „Cup“ ?

„Cup“

„Pedestal“

Handlungsplan (action plan) makes a thing become an object

„Cup“ „Pedestal“

Handlungsplan (action plan) makes a thing become an object

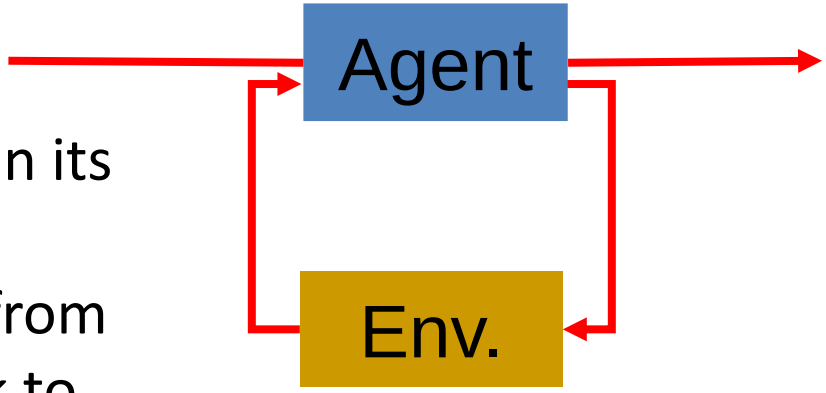
Objects and Actions are inseparably intertwined!

Wörgötter et al (2009). RAS, 57(4):420-432.

Krüger, N., Piater, J., Geib, C., Petrick, R., Steedman, M., Wörgötter, F., Ude, A., Asfour, T., Kraft, D., Omrcen, D., Agostini, A. and Dillmann, R. (2011). Object-Action Complexes: Grounded Abstractions of Sensorimotor Processes, RAS, 59(10), 740-757

Objects and Actions are inseparably intertwined!

To “understand” an object an agent must be able to act (or simulate an action mentally)



For this an agent has to be **situated** in its world. There must be a **closed loop** where information from the world (from the agent’s own actions) comes back to the agent.

Some notes:
Conventional AI systems where not situated (if anything their feedback was provided by the programmer or the user)

The environment defines the body (**embodiment**). The **interaction** with the environment (**situatedness**) allows the development and the grounding of symbols.

There are non-embodied (or only mildly embodied) systems that are indeed situated (that interact with the word) and that do show very complex signs of intelligence!

Swarms, Societies (also human societies) are such systems.

How can we arrive at a representation that captures objects in an action context?

How can we capture actions??

This can only be achieved in a procedural way!

Problems and Questions

Hardware

Substrate

Cartesian (knowledge w.o. substrate) versus Embodied (knowledge needs a body)

Data Formats

Computation and storage:

Symbolic (Computer programs) versus Sub-symbolic (Connectionism)

Software

Processes !!

Knowledge acquisition:

Outside-In (exploration based) versus Inside-Out (generative, experience based)

The Weird Stuff

Philosophy of Mind

Hypotheses: Strong AI versus Weak AI

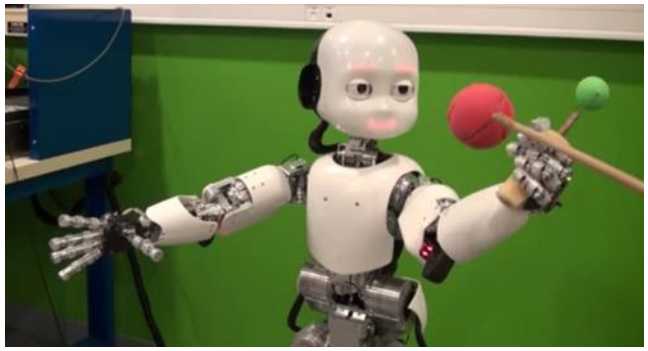
Strong AI Hypothesis: A computer which behaves as intelligently as a person must also necessarily have a mind and consciousness (whatever that is…..Monism, Dualism, etc.).

How can we arrive at a representation that captures objects in an action context?

How can we capture actions??

This can only be achieved in a procedural way!

Very many approaches exist to create learning “robot babies” that acquire knowledge through exploring the world. This is very slow and tedious.



But: How can we arrive at a fast generative process for knowledge acquisition? (Inside out!)

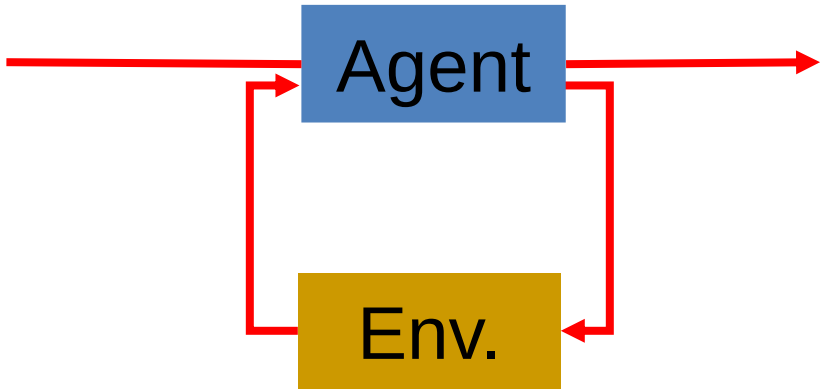
Software

Knowledge acquisition:

Processes !!

Outside-In (exploration based) versus Inside-Out (generative, experience based)

To “understand” an object an agent must be able to act (or simulate an action mentally)



We need a representation that is fundamentally procedural !

How to capture actions & derive generative processes from this?

What's the problem here ?

Example: Understand how to „Make a sandwich“

Objects involved

















How to capture actions & derive generative processes from this?

Example part 1: Understand how to „Make a sandwich“

Objects involved + Action

Breads

Hand Spreads

Tools

Cheese or Salami

How to capture actions & derive generative processes from this?

Example: Understand how to „Make a sandwich“

Objects involved + Action

Breads

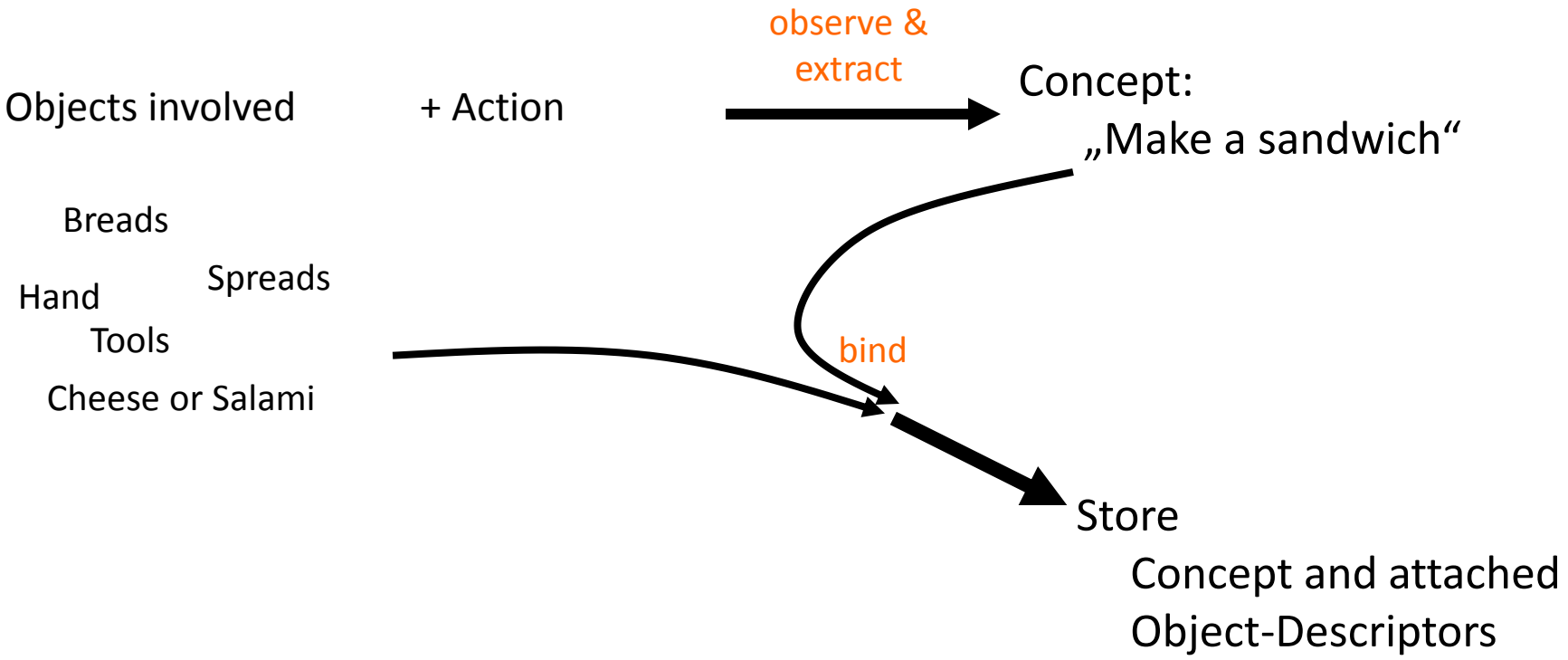
Hand Spreads

Tools

Cheese or Salami

How to capture actions & derive generative processes from this?

Example: Understand how to „Make a sandwich“



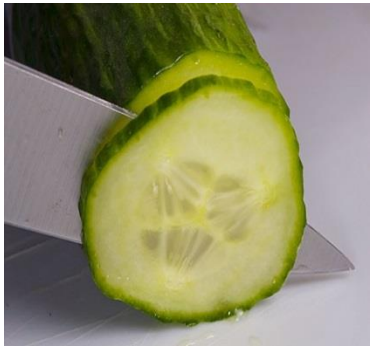
Actions: A grammatical, sequential view!

Actions: A grammatical, sequential view!

- Hand-only action
  - Rearrange : push, poke, flick, Stir

Actions: A grammatical, sequential view!

- Hand-only action
  - Destroy : cut, chop, pinch



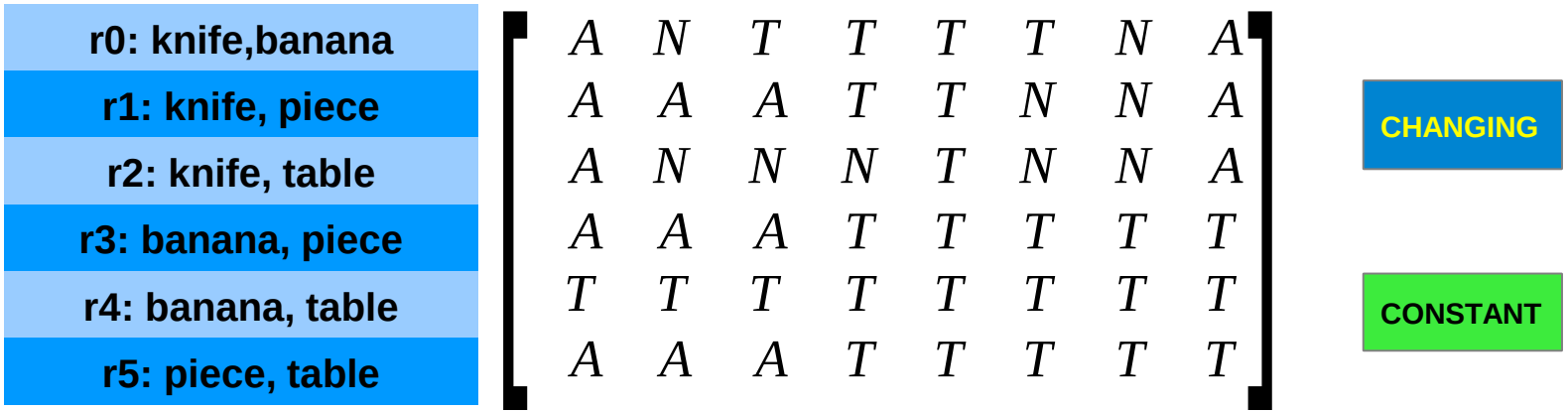
Wörgötter, F., Aksoy, E.E., Krüger, N., Piater J., Ude, A. and Tamosiunaite M. (2013). A Simple Ontology of Manipulations: Towards representations for manipulation actions in robotics. IEEE Transact. Autonomous Mental Development (TAMD) 5(2), 117-134.

Semantic Event Chain

- Graphs → Matrix

Hand , object $\begin{bmatrix} N & T & N \end{bmatrix}$

Example: Cutting a Banana



Here we have arrived at a procedural representation of a simple action free from the actual objects on which it has been performed.

So we may have a representation, but where is the generative process?

Software

Knowledge acquisition:

Outside-In (exploration based) versus Inside-Out (generative, experience based)

Processes !!

?

Piaget's Accommodation and Assimilation

Assimilation: Entering new entities into existing schemas.

Accommodation: Storing new schemas.

Jean Piaget

1896 – 1980

Schema for Mammals: Mammals have 4 legs.

Schema: Mammal_4: Dog

Assimilation:
A cat has 4 legs: This fits to the schema and, thus, it is a mammal.

Schema extension: Mammal_4:Dog, Cat

Dolphins have no legs but they are mammals.

Accommodation:
Mammals can have no legs (new schema)

Two Schemas: Mammal_4
: Mammal_0

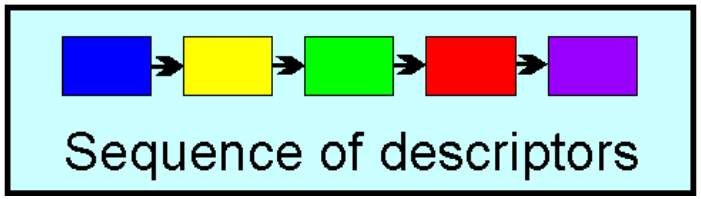
Robotics Example:
Cutting Chopping Stirring


Frame no: 2393
Frame no: 2442
Frame no: 2470
Frame no: 2500
Frame no: 3031
Frame no: 1212
Frame no: 1234
Frame no: 1302
Frame no: 1458
Frame no: 1547
Frame no: 1585
Frame no: 1337
Frame no: 1384
Frame no: 1514
Frame no: 1525
Frame no: 1779
Frame no: 1816

Piaget's Accommodation and Assimilation

In Memory:

Action Schema for Action α



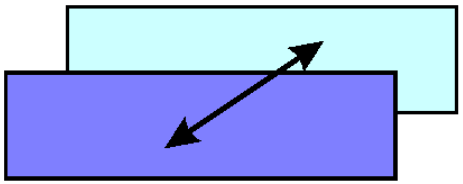
„Cutting“

A	N	T	T	T	T	N	A
A	A	A	T	T	N	N	A
A	N	N	N	T	N	N	A
A	A	A	T	T	T	T	T
T	T	T	T	T	T	T	T
A	A	A	T	T	T	T	T

Semantic Comparison

at the level of the "whole" Action

Action α

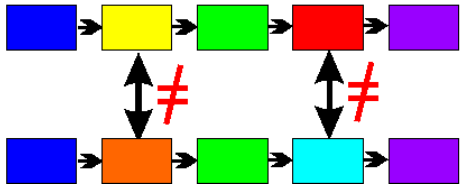


Action β

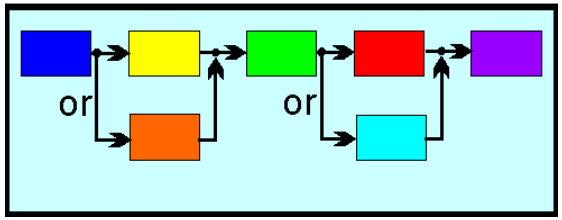
Yes

Syntactic Comparison

"inside" each Action



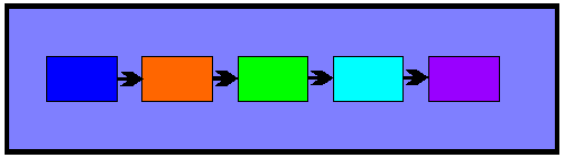
Assimilation



„Chopping“

No

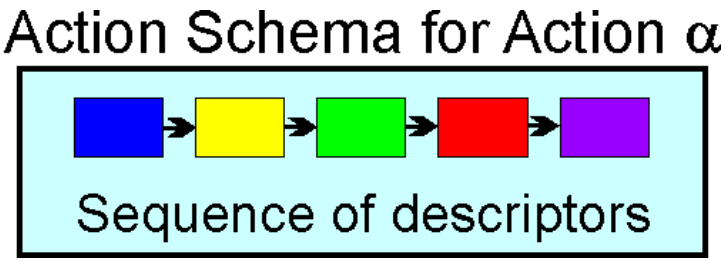
Accommodation



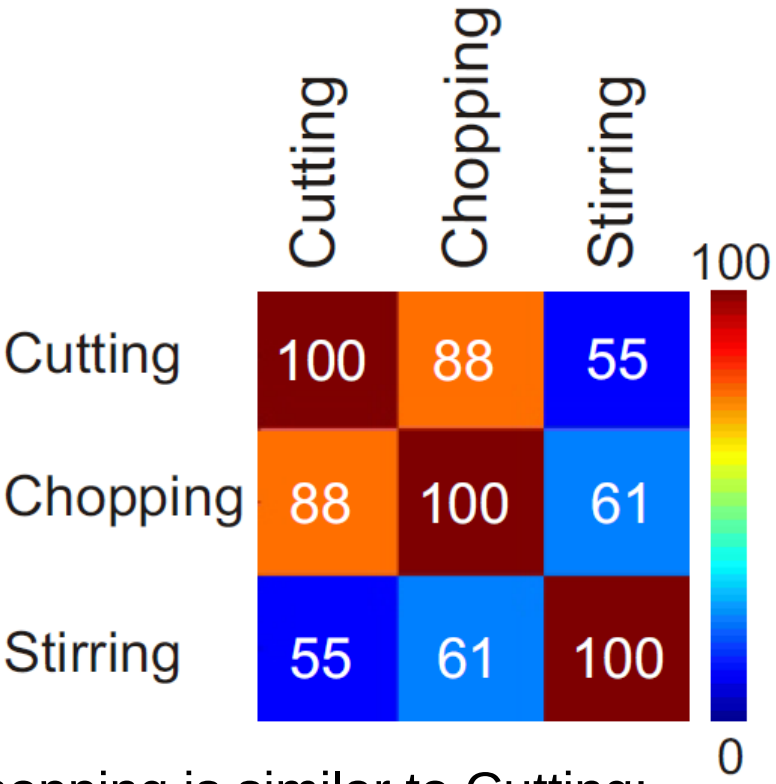
New Action Schema (Action β)

„Stirring“

In Memory:



„Cutting“



Stirring is not similar to Cutting:
Store new schema

Accomodation: Stirring

Chopping is similar to Cutting:
Perform syntactic comparison

Assimilation: Cutting: Cutting-proper, Chopping

Assimilation

Stored Experience

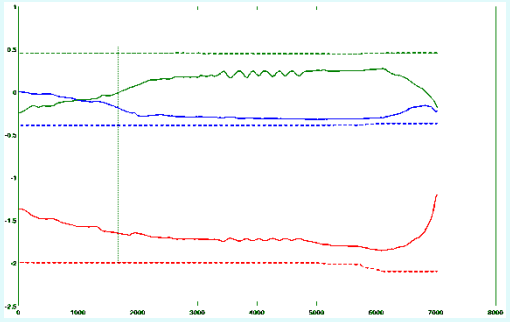
Cutting

table, knife	-1	0	1	1	1	1	1	1	1	1	0	-1
table, hand	-1	-1	-1	0	0	0	1	1	1	0	-1	-1	-1
table, new piece	-1	-1	-1	-1	1	1	1	1	1	1	1	1	1
cucumber, knife	-1	0	0	0	1	1	1	0	0	0	0	0	-1
cucumber, new piece	-1	-1	-1	-1	-1	0	0	1	0	0	0	0	0
knife, hand	-1	-1	-1	1	1	1	1	1	1	1	-1	-1	-1
knife, new piece	-1	-1	-1	-1	-1	1	1	0	0	0	0	0	-1

Objects

Cucumber Knife Hand

Trajectories



Novel Observation

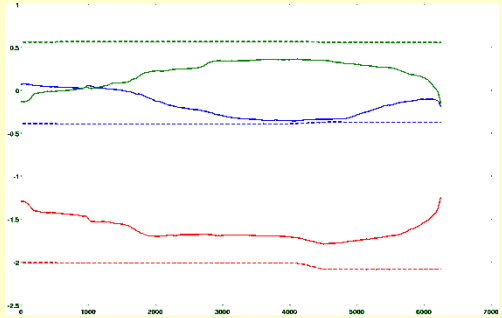
Chopping

table, cleaver	-1	0	0	1	1	1	1	1	1	1	1	0	-1
table, hand	-1	-1	0	0	0	1	1	1	0	0	-1	-1	-1
table, new piece	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1
carrot, knife	-1	0	0	0	1	1	1	0	0	0	0	0	-1
cleaver, hand	-1	-1	-1	1	1	1	1	1	1	1	-1	-1	-1
cleaver, new piece	-1	-1	-1	-1	-1	-1	1	1	1	0	0	0	-1

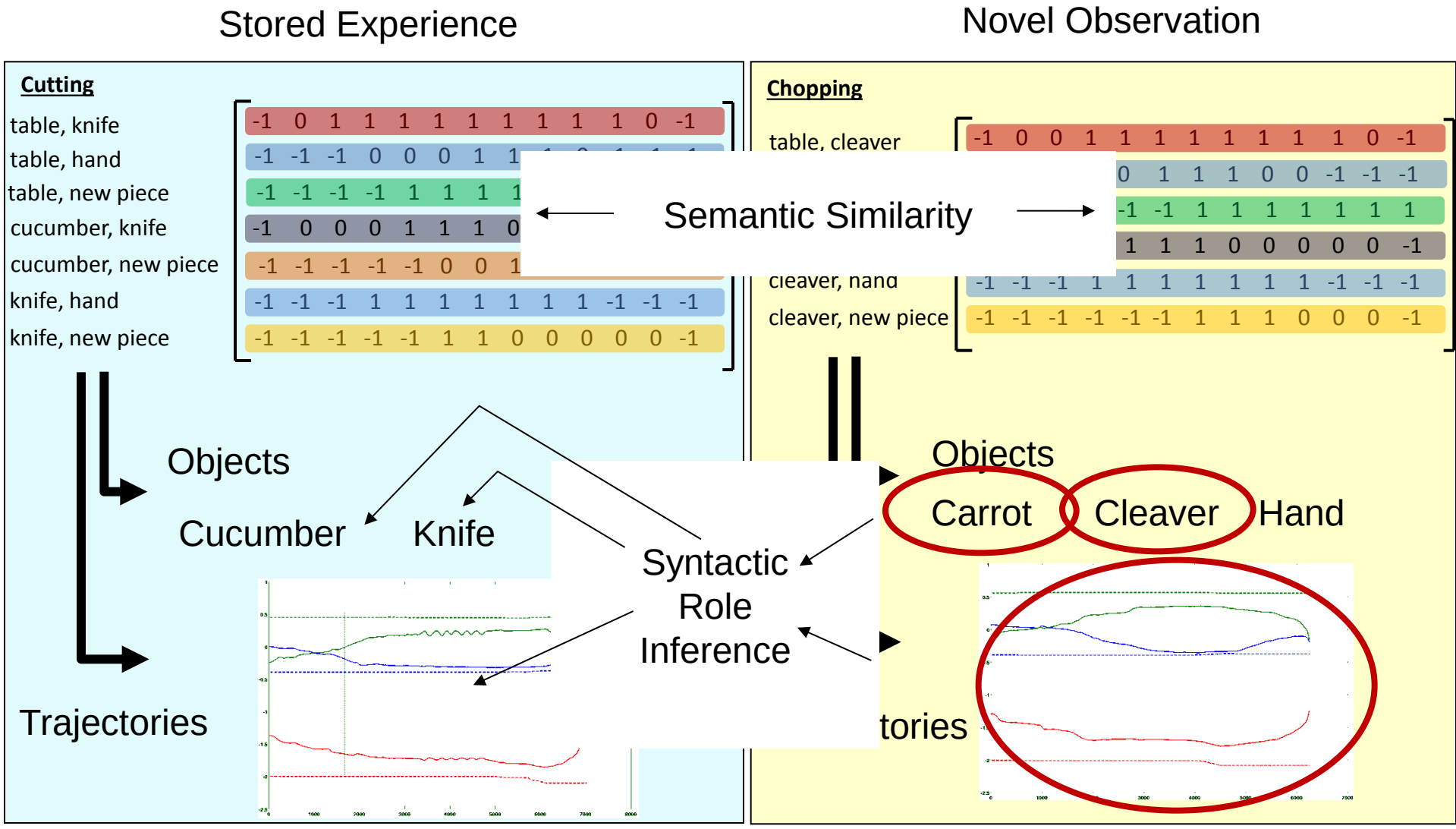
Objects

Carrot Cleaver Hand

Trajectories



Assimilation (cont.)



The Broader Perspectives

- Event Chains are an (object-free) procedural representation of actions based on observable (grounded!) events.
- Piagetian or other similar processes allow reasoning and generative knowledge acquisition
- The Xperience Project subsumes such processes under the term “Structural Bootstrapping” (Inside-Out!)

More on Symbols (Language)

Grammar:

Subject + Verb + Direct Object + Indirect Object

Manipulation:

Manipulator + Action + Primary Object + Secondary Object

Example:

The hand puts a cup on top of a box

SEC:

Subject, Dir. Obj., Indir. Obj.,

A	N	T	T	N	A
N	N	N	T	T	T

Verb (pose, speed, etc.)

More on Symbols (Language)

Grammar:

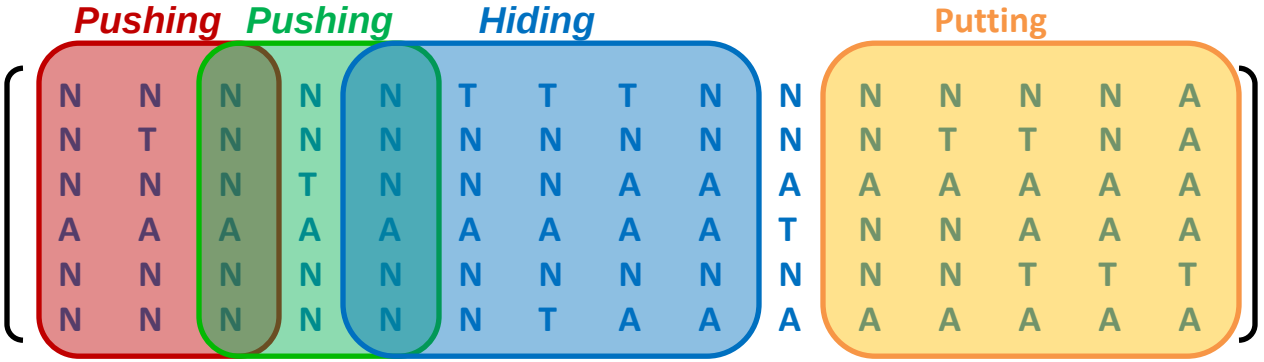
Subject + Verb + Direct Object + Indirect Object

Example:

The hand first pushes a box then a ball.
The hand afterwards hides the ball
with a bucket and puts the box
on top of the bucket.

SEC:

hand , bucket
hand , box
hand , ball
hand , noise
box , bucket
ball , bucket



Something **very strange** has happened here!
Through naked observation of signals (vision), events are recorded (touching, etc.).
These events bridge the gap between signals and symbols (language)

Some final notes:
The agent does not have to have “our” linguistic understanding of the (by us) uttered sentences.

As long as it can in a self-consistent way observe and extract the Event Chains (encoded in any way it pleases) and reproduce the actions by execution it has arrived at a fully grounded and very general understanding of these actions.

Does it need to have a mind to understand “more” or to understand this “really”?

Or is this all it takes to have a (limited) mind?

The Weird Stuff

Philosophy of Mind

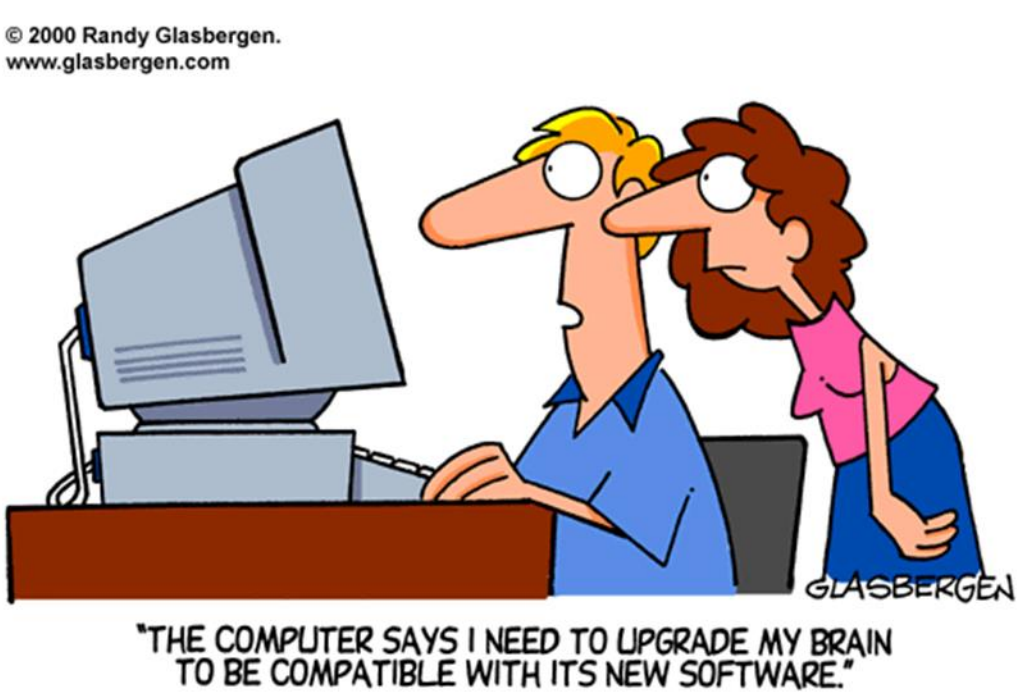
Hypotheses: Strong AI versus Weak AI

Strong AI Hypothesis: A computer which behaves as intelligently as a person must also necessarily have a mind and consciousness (whatever that is…..Monism, Dualism, etc.).

This topic must be deferred to →

Acknowledgements

Bernd Porr (Glasgow)
Nobert Krüger (Odense)
Eren Aksoy (Gö)
Minija Tamosiunaite (Gö)
And many others from
PACO-Plus and Xperience



EU-FP7

PACO+, Xperience, IntellAct, ACAT,

BMBF

BCCN

Some remarks on Philosophy of Mind (the “Weird Stuff”)

Qualia (/ˈkwɑːliə/ or /ˈkweɪliə/; singular form: quale (Latin pronunciation: [ˈkwaːle]) is a term used in philosophy to refer to individual instances of subjective, conscious experience. The term derives from a Latin word meaning for "what sort" or "what kind." Examples of qualia are the pain of a headache, the taste of wine, or the perceived redness of an evening sky.

Daniel Dennett identifies four properties that are commonly ascribed to qualia. According to these, qualia are:

1. ineffable; that is, they cannot be communicated, or apprehended by any other means than direct experience.
2. intrinsic; that is, they are non-relational properties, which do not change depending on the experience's relation to other things.
3. private; that is, all interpersonal comparisons of qualia are systematically impossible.
4. directly or immediately apprehensible in consciousness; that is, to experience a quale is to know one experiences a quale, and to know all there is to know about that quale.

If qualia of this sort exist, then a normally sighted person who sees red would be unable to describe the experience of this perception in such a way that a listener who has never experienced color will be able to know everything there is to know about that experience

Arguments in favor

In an article "Epiphenomenal Qualia" (1982),[13] Frank Jackson offers what he calls the "knowledge argument" for qualia. One example runs as follows:

Mary the colour scientist knows all the physical facts about colour, including every physical fact about the experience of colour in other people, from the behavior a particular colour is likely to elicit to the specific sequence of neurological firings that register that a colour has been seen. However, she has been confined from birth to a room that is black and white, and is only allowed to observe the outside world through a black and white monitor. When she is allowed to leave the room, it must be admitted that she learns something about the colour red the first time she sees it — specifically, she learns what it is like to see that colour.

This thought experiment has two purposes. First, it is intended to show that qualia exist. If we agree with the thought experiment, we believe that Mary gains something after she leaves the room—that she acquires knowledge of a particular thing that she did not possess before. That knowledge, Jackson argues, is knowledge of the quale that corresponds to the experience of seeing red, and it must thus be conceded that qualia are real properties, since there is a difference between a person who has access to a particular quale and one who does not.

The second purpose of this argument is to refute the physicalist account of the mind. Specifically, the knowledge argument is an attack on the physicalist claim about the completeness of physical truths. The challenge posed to physicalism by the knowledge argument runs as follows:

1. Before her release, Mary was in possession of all the physical information about color experiences of other people.
2. After her release, Mary learns something about the color experiences of other people.
3. Therefore,
4. Before her release, Mary was not in possession of all the information about other people's color experiences, even though she was in possession of all the physical information.

Therefore,

1. There are truths about other people's color experience that are not physical.
2. Therefore,
3. Physicalism is false.

Refuting this

Dennett also has a response to the "Mary the color scientist" thought experiment. He argues that Mary would not, in fact, learn something new if she stepped out of her black and white room to see the color red. Dennett asserts that if she already truly knew "everything about color," that knowledge would include a deep understanding of why and how human neurology causes us to sense the "quale" of color. Mary would therefore already know exactly what to expect of seeing red, before ever leaving the room. Dennett argues that the misleading aspect of the story is that Mary is supposed to not merely be knowledgeable about color but to actually know all the physical facts about it, which would be a knowledge so deep that it exceeds what can be imagined, and twists our intuitions.

If Mary really does know everything physical there is to know about the experience of color, then this effectively grants her almost omniscient powers of knowledge. Using this, she will be able to deduce her own reaction, and figure out exactly what the experience of seeing red will feel like.

Dennett finds that many people find it difficult to see this, so he uses the case of **RoboMary** to further illustrate what it would be like for Mary to possess such a vast knowledge of the physical workings of the human brain and color vision. RoboMary is an intelligent robot who, instead of the ordinary color camera-eyes, has a software lock such that she is only able to perceive black and white and shades in-between.

RoboMary can examine the computer brain of similar non-color-locked robots when they look at a red tomato, and see exactly how they react and what kinds of impulses occur. RoboMary can also construct a simulation of her own brain, unlock the simulation's color-lock and, with reference to the other robots, simulate exactly how this simulation of herself reacts to seeing a red tomato. RoboMary naturally has control over all of her internal states except for the color-lock. With the knowledge of her simulation's internal states upon seeing a red tomato, RoboMary can put her own internal states directly into the states they would be in upon seeing a red tomato. In this way, without ever seeing a red tomato through her cameras, she will know exactly what it is like to see a red tomato.

Dennett uses this example to show us that Mary's all-encompassing physical knowledge makes her own internal states as transparent as those of a robot or computer, and it is almost straightforward for her to figure out exactly how it feels to see red.

Another argument pro “Qualia”: ThePhilosophical Zombies

It is **logically conceivable** that there could be physical duplicates of people, called "zombies," without any qualia at all. These "zombies" would demonstrate outward behavior precisely similar to that of a normal human, but would not have a subjective phenomenology.
(It is worth noting that a necessary condition for the possibility of philosophical zombies is that there be no specific part or parts of the brain that directly give rise to qualia—the zombie can only exist if subjective consciousness is causally separate from the physical brain.)

The arguments used to refute this a build on the conjecture that philosophical zombies cannot exist (not even in a “logical sense”! See further slides.

By the way: All philosophers agree that such zombies would not exist in a physical way anyways.

But……. **What about Robots??** Would they not fall into the “Zombie” Category (according to those philosophers)!!?? (Admittedly robots are still too simple to be considered real Zombies, but maybe in 20,30,50, 100…. Years.

Minsky says: Now, a philosophical dualist might then complain: "You've described how hurting affects your mind — but you still can't express how hurting feels." This, I maintain, is a huge mistake — that attempt to reify 'feeling' as an independent entity, with an essence that's indescribable. As I see it, feelings are not strange alien things. It is precisely those cognitive changes themselves that constitute what 'hurting' is — and this also includes all those clumsy attempts to represent and summarize those changes. The big mistake comes from looking for some single, simple, 'essence' of hurting, rather than recognizing that this is the **word** we use for **complex rearrangement** of our disposition of resources

Our own thinking:

There is a certain lack of arguments w.r.t. “emergence”. Minsky touches upon this. Maybe qualia are just an **emergent phenomenon**?? Its capturing into a single expression is just not rich enough to convey it to others (Minsky’s argument)!

Also: There is a certain link to the problem of **non-declarative knowledge (**motor skills like skiing). Also on this end there is no way of directly conveying this type of knowledge to “someone else” (by language or by **other symbolic means**). This is different from declarative knowledge like maths.

Hence one does not have to go done the whole strange route to qualia and mind to find a related problem, which is closer to an “undeniable” physicalist explanation… Do you need an non-material mind to learn and do skiing???????????

More On Philosophical Zombies

A **philosophical zombie** or p-zombie in the philosophy of mind and perception is a hypothetical being **that is indistinguishable from a normal human being except in that it lacks conscious experience**, qualia, or sentience.[1] When a zombie is poked with a sharp object, for example, it does not feel any pain though it behaves exactly as if it does feel pain (it may say "ouch" and recoil from the stimulus, or tell us that it is in intense pain).

Though philosophical zombies are widely used in thought experiments, the detailed articulation of the concept is not always the same. P-zombies **were introduced primarily to argue against specific types of physicalism** such as behaviorism, according to which mental states exist solely as behavior: belief, desire, thought, consciousness, and so on, are simply certain kinds of behavior or tendencies towards behaviors. **A p-zombie that is behaviorally indistinguishable from a normal human being but lacks conscious experiences is therefore not logically possible according to the behaviorist, so an appeal to the *logical possibility* of a p-zombie furnishes an argument that behaviorism is false.** Proponents of zombie arguments generally accept that p-zombies are not physically possible, while opponents necessarily deny that they are even logically possible.

Artificial intelligence researcher Marvin Minsky sees the argument as circular. The proposition of the possibility of something physically identical to a human but without subjective experience assumes that the physical characteristics of humans are not what produces those experiences, which is exactly what the argument was claiming to prove.[13]

See Chalmers "The conscious mind" (1996) for opposition see Dennett

- The **frame problem** is that specifying only which conditions are changed by the actions do not allow, in logic, to conclude that all other conditions are not changed. For example, if the action executed at time 0 is that of opening a door, you cannot conclude that the light has not changed, too.

  More philosophically: Is it possible, in principle, to limit the scope of the reasoning required to derive the consequences of an action? And, more generally, **how do we account for our apparent ability to make decisions on the basis only of what is relevant to an ongoing situation without having explicitly to consider all that is not relevant?**

- Related are: The **qualification problem** is concerned with the impossibility of listing all the preconditions required for a real-world action to have its intended effect. It might be posed as **how to deal with the things that prevent me from achieving my intended result**.

  McCarthy said: "The successful use of a boat to cross a river requires, if the boat is a rowboat, that the oars and rowlocks be present and unbroken, and that they fit each other. Many other qualifications can be added, making the rules for using a rowboat almost impossible to apply, and yet anyone will still be able to think of additional requirements not yet stated."

- And the: The **ramification problem** is concerned with the indirect consequences of an action. It might also be posed as **how to represent what happens implicitly due to an action** or how to control the secondary and tertiary effects of an action.

Due to these and other problems (Perceptron book of Minsky and Papert) AI research came to a **virtual standstill between 1970 and 1980**.

Next successes (AI revival 1980-1990):

- Expert Systems: Limited domain, expert knowledge (no need for vast "common sense knowledge", constrained reasoning. → Specification!
- "Cyc" Data base: First attempt to store common sense knowledge in a large scale.
- Connectionisms revived:

  John Hopfield (1982), a novel type of powerful network with math proof!
  Paul Werbos, David Rumelhart "error backpropagation" to train ANNs
  Appearance of "Parallel Distributed Processing" book (1986), David Rumelhart & James McClelland.
  Commercial impact like OCR, speech recognition and process control (around 1990).

Next failures (Systems did not scale up!):

- Expert System proved to be brittle (non-robust, giving "strange answers")
- The "irrelevant" philosophical problems (frame problem, qualification problem) killed larger AI system.
- Several promises (like "make a conversation with a machine) had not been achieved.

Again AI research came to a halt!