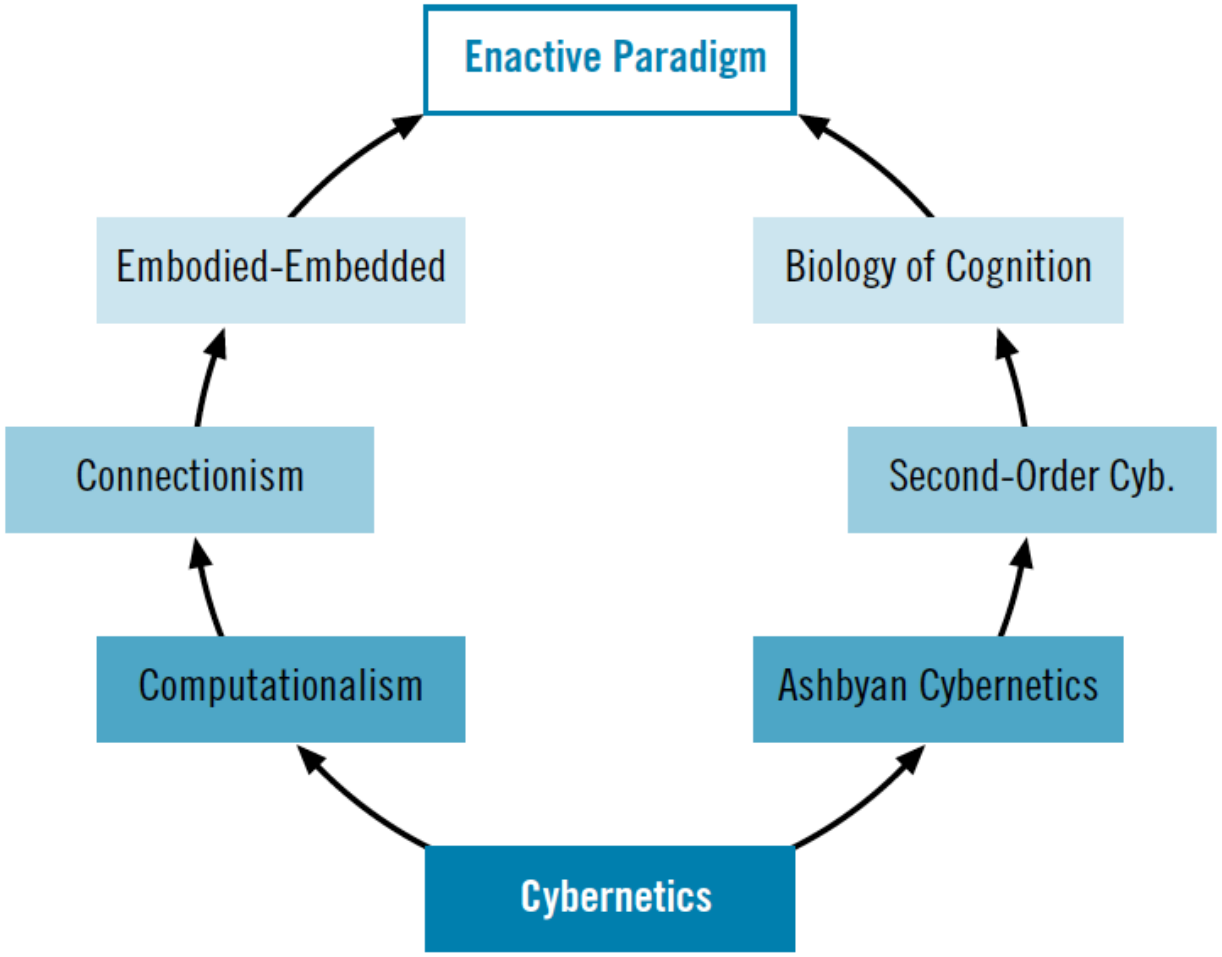
OVERCOMING THE LIMITS OF AI BY EMPOWERING THE HUMAN MIND

Tom Froese

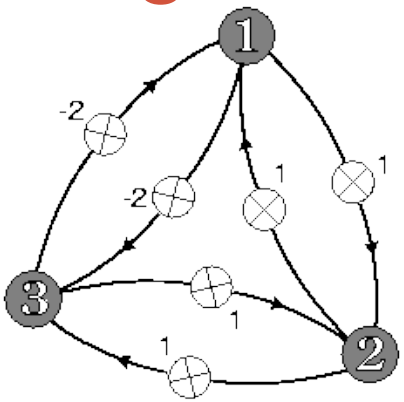
Embodied Cognitive Science Unit
Okinawa Institute of Science and Technology Graduate University

History of enactive cognitive science



Froese (2010)

A brief history of AI and cognitive robotics



Symbolic, sub-symbolic, embodied, ...

Internalist to relational theory of mind

The problem of scaling up

Artificial Intelligence 47 (1991) 161–184
Elsevier

Today the earwig, tomorrow man?

David Kirsh
Department of Cognitive Science C-015, University of California, San Diego, La Jolla, CA 92093, USA

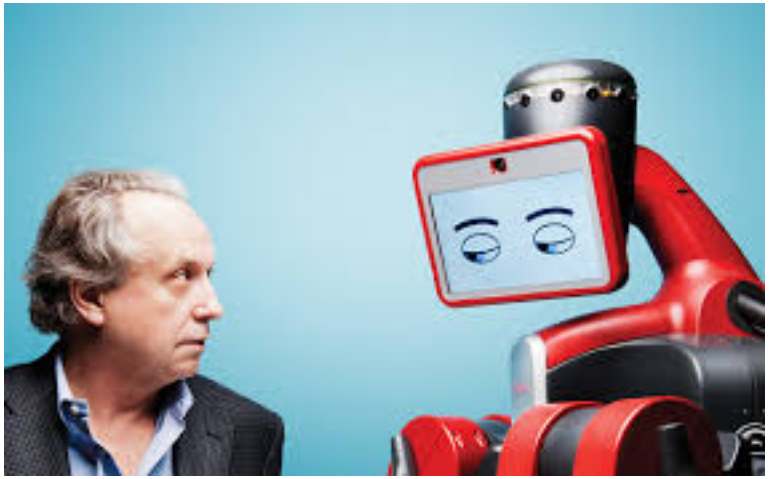
ELSEVIER

Robotics and Autonomous Systems 20 (1997) 291–304

Robotics and Autonomous Systems

From earwigs to humans

Rodney A. Brooks [1]
MIT Artificial Intelligence Laboratory, 545 Technology Square, Cambridge, MA 02139, USA



Good Old Fashioned AI

Deep neural networks

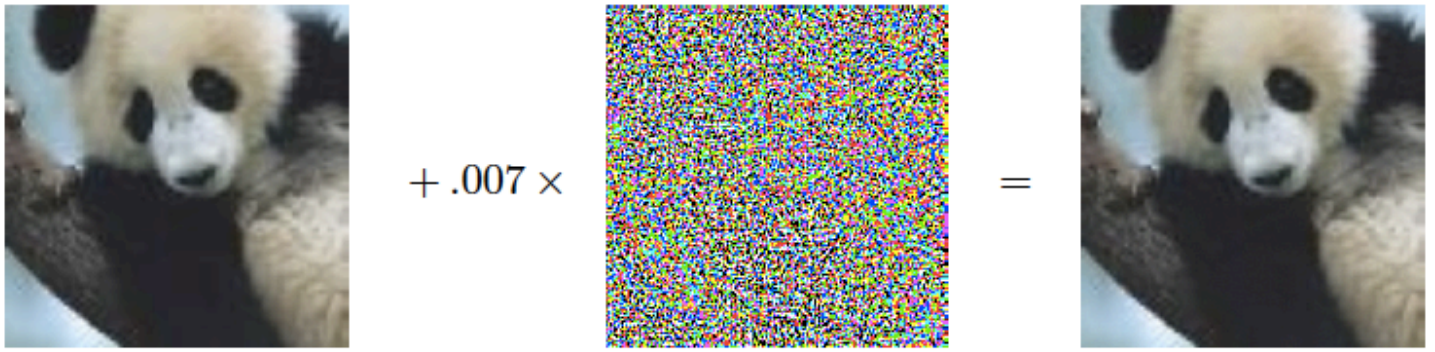


Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

Goodfellow et al. (2015)

Uber self-driving cars no more…?

AARIAN MARSHALL AND ALEX DAVIES TRANSPORTATION 05.24.18 03:38 PM

UBER'S SELF-DRIVING CAR SAW THE WOMAN IT KILLED, REPORT SAYS

The National Transportation Safety Board says Uber's self-driving car had trouble identifying Elaine Herzberg as a human, and that it couldn't hit the brakes to avoid hitting her.

TRIPPLAAR KRISTOFFER/SIPA VIA AP IMAGES

The report says that the Uber vehicle, a modified Volvo XC90 SUV, had been in autonomous mode for 19 minutes and was driving at about 40 mph when it hit 49-year-old Elaine Herzberg as she was walking her bike across the street. The car's radar and lidar sensors detected Herzberg about six seconds before the crash—first identifying her as an unknown object, then as a vehicle, and then as a bicycle, each time adjusting its expectations for her path of travel.

About a second before impact, the report says "the self-driving system determined that an emergency braking maneuver was needed to mitigate a collision." Uber, however, does not allow its system to make emergency braking maneuvers on its own. Rather than risk "erratic vehicle behavior"—like slamming on the brakes or swerving to avoid a plastic bag—Uber relies on its human operator to watch the road and take control when trouble arises.

The money keeps flowing!

Paul Allen Wants to Teach Machines Common Sense

"To make real progress in A.I., we have to overcome the big challenges in the area of common sense," said Paul Allen, who founded Microsoft in the 1970s with Bill Gates. Béatrice de Géa for The New York Times

By Cade Metz

Feb. 28, 2018

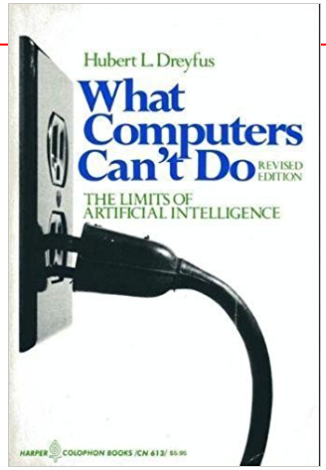
SAN FRANCISCO — Microsoft's co-founder Paul Allen said Wednesday that he was pumping an additional $125 million into his nonprofit computer research lab for an ambitious new effort to teach machines "common sense."

In the mid-1980s, Doug Lenat, a former Stanford University professor, with backing from the government and several of the country's largest tech companies, started a project called Cyc. He and his team of researchers worked to codify all the simple truths that we learn as children, from "you can't be in two places at the same time" to "when drinking from a cup, hold the open end up."

Thirty years later, Mr. Lenat and his team are still at work on this "common sense engine" — with no end in sight.

Mr. Allen helped fund Cyc, and he believes it is time to take a fresh approach, he said, because modern technologies make it easier to build this kind of system.

Mr. Lenat welcomed the new project. But he also warned of challenges: Cyc has burned through hundreds of millions of dollars in funding, running into countless problems that were not evident when the project began. He called them "buzz saws."



Hume's fact-value gap

- Hume's law:
  - There is a gap between is-statements and ought-statements. No amount of descriptive facts can force a **normative** choice.
  - "Where a passion is neither founded on false suppositions, nor chooses means insufficient for the end, the understanding can neither justify nor condemn it.
  - **'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger.**" (§2.3.3.6)

Hume, D. (1739–1740a). *A Treatise of Human Nature*.

What is normativity?

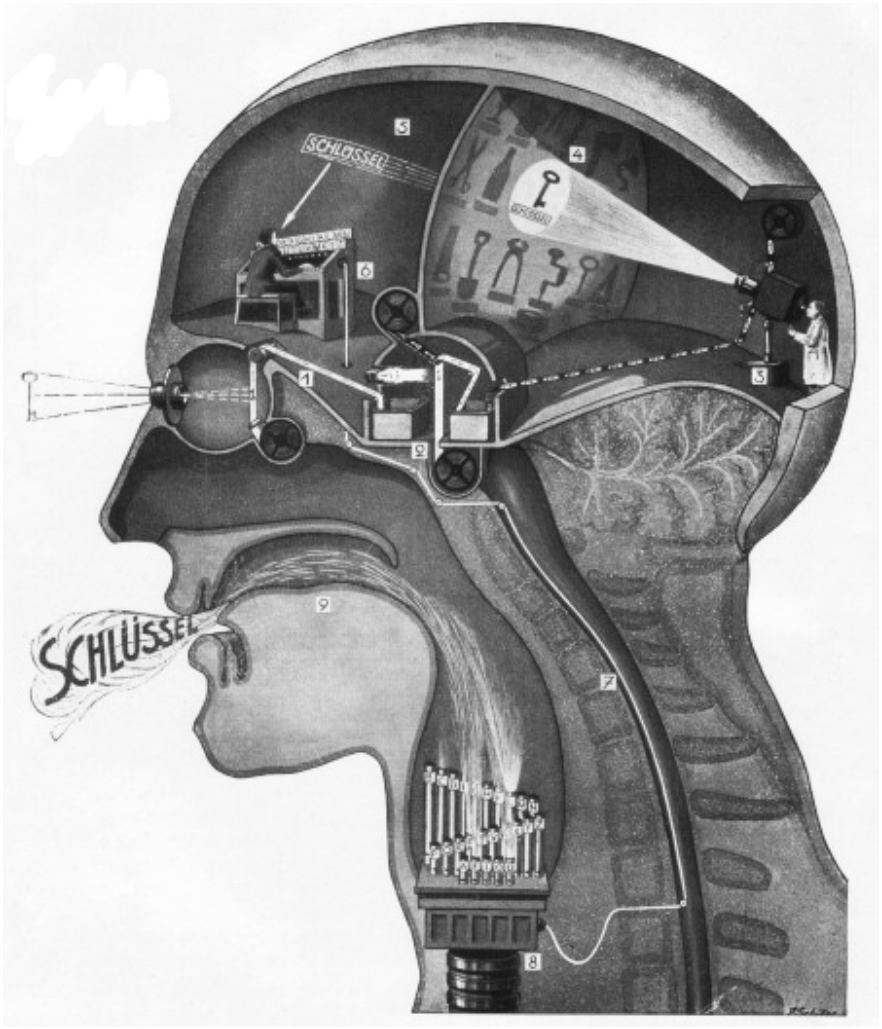
- An activity is more than mere movement (or: movement is less than an activity – it lacks or misses something).
  - An activity is someone **doing** something (e.g. you are listening).
  - It is **not** only a process that is **happening** (e.g. the earth moves).
  - Activity is done **for** or **in order to** realize something (a purpose, goal, desire, intention, etc.).
- Therefore, activity is **normative** because it can succeed or fail with respect to certain evaluative criteria.

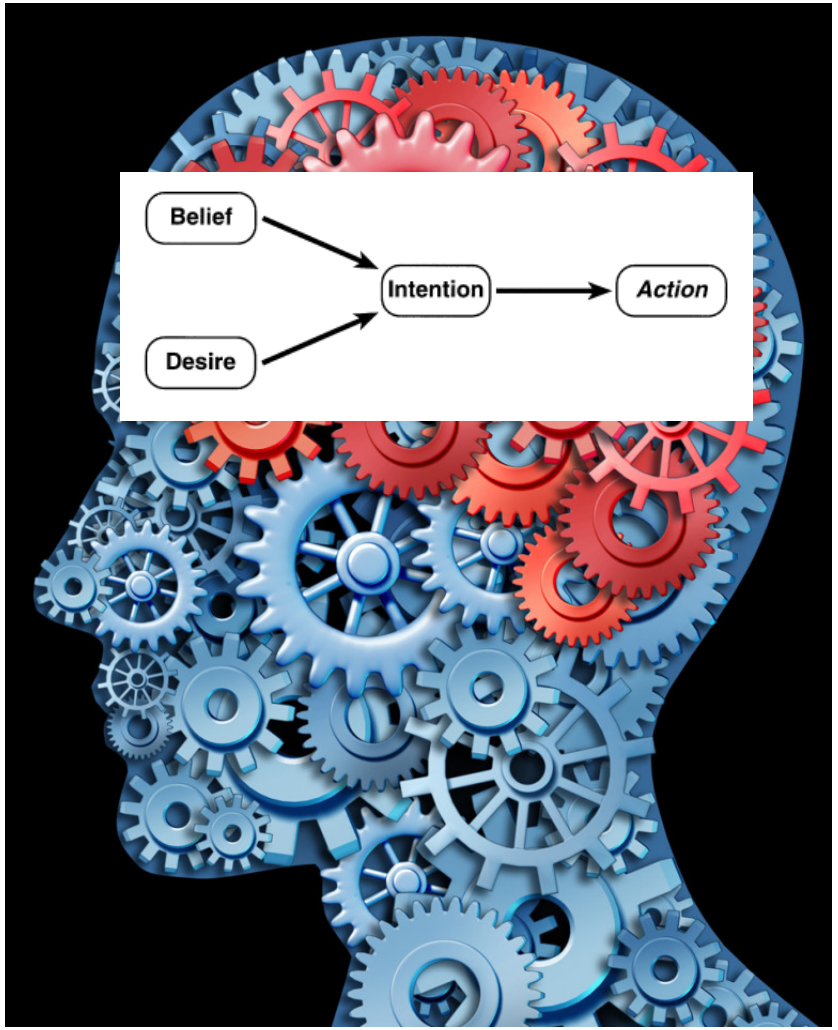
The origins of “yum and yuck”

- “Once there is an autonomous agent, there is a **semantics** from its privileged point of view. The incoming molecule is ‘yuck’ **or** ‘yum’. …
- I think that from the autonomous agent’s perspective, yuck or yum is primary, unavoidable, and of the **deepest importance to that agent**. …
- the rudiments of **value** are present once autonomous agents are around.” (p. 111-117)

Kauffman, S. (2000). *Investigations*. New York, NY: Oxford University Press

Normativity v1: GOFAI



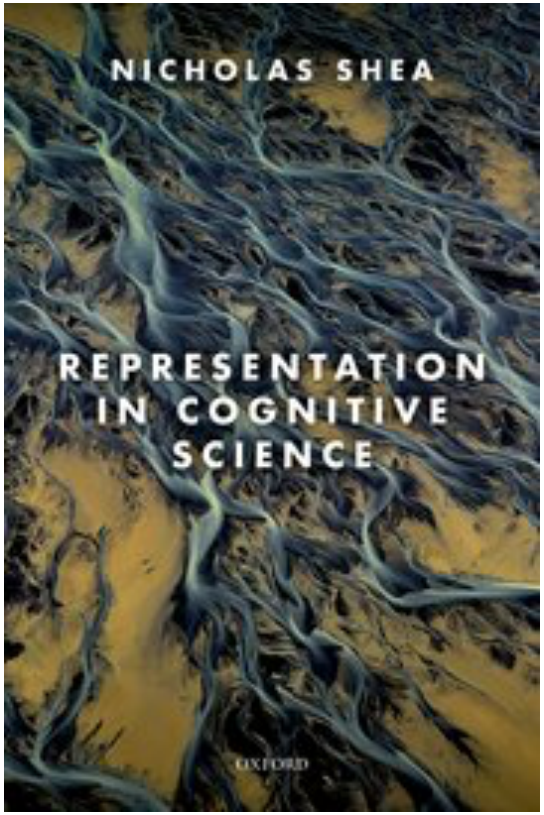


Searle's (1980) "Chinese room" argument

A problem of normativity

- Harnad's (1990) "symbol grounding problem"
- McCarthy and Hayes' (1969) "frame problem", and more general versions by Dennett (1984) and Wheeler (2005)
- Searle's (1990) "Chinese Room" argument
- Dreyfus' (1972, 1992) problems of meaning and of commonsense knowledge
- In general, the root problem of GOFAI is an **in principle** failure to logically determine what **ought** to be meant or done given what is **fact** and what is **true** (Froese 2009).

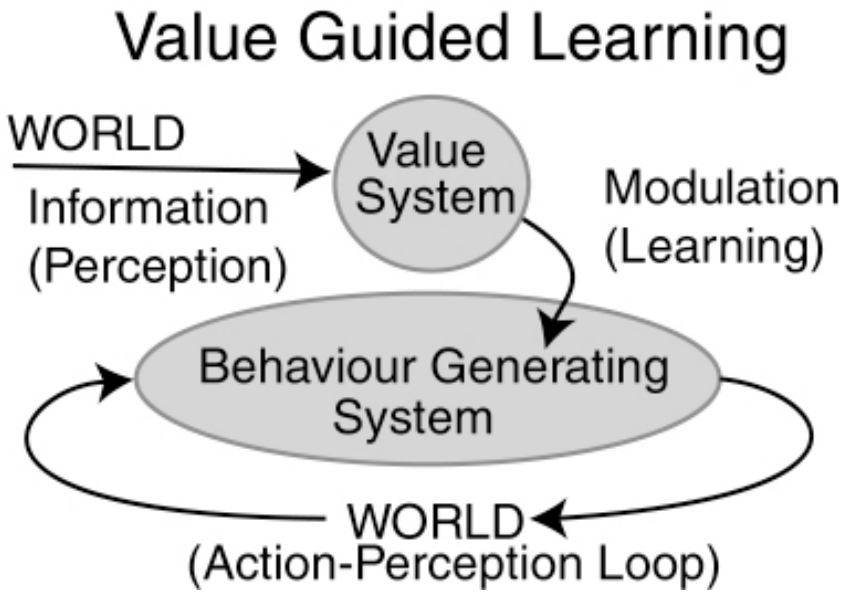
The problems of representation are still with us!



- "we can make a machine whose manipulations obey logical rules and so preserve truth.
- But **we don't yet have a clear idea of how representations could get meanings, when the meaning does not derive from the understanding of an external interpreter."**
  - (Shea 2018, p. 4)

Normativity v2.0: Embodied AI

Principle	Name
Types of agents of interest, ecological niche and tasks	
1	The "complete agents" principle
2	The "ecological niche" principle
Morphology, architecture, mechanism	
3	The principle of parallel, loosely coupled processes (the "anti - homunculus" principle)
4	The "value" principle
5	The principle of sensory-motor coordination
6	The principle of "ecological balance"
7	The principle of "cheap designs"
Strategies, heuristics, stances, metaphors	
8	"Frane-of-reference" principle
9	"Constraints" principles
10	Compliance with principles
	etc.



Pfeifer, R. (1996). Building "fungus eaters:" Design principles of autonomous agents. In P. Maes et al. (Eds.), *From Animals to Animats 4* (pp. 3-12). Cambridge, MA: MIT Press.

Di Paolo, E. A. et al. (2010). Horizons for the enactive mind: Values, social interaction, and play. In J. Stewart et al. (Eds.), *Enaction* (pp. 33-87). Cambridge, MA: MIT Press

The failure of embodied cognition?

- "The current flourishing of embodied and situated approaches to AI, cognitive science and robotics has shown that the arguments from that period [i.e. the 1990s] were indeed convincing to many,

- but time and reflection has in fact cast **doubt** on whether they were right.

- This is precisely the situation that most calls out for **philosophical reflection**."

M.L. Anderson, Strike while the iron is, *Artificial Intelligence* **170** (18) (2006) 1213–1217

A problem of normativity, again

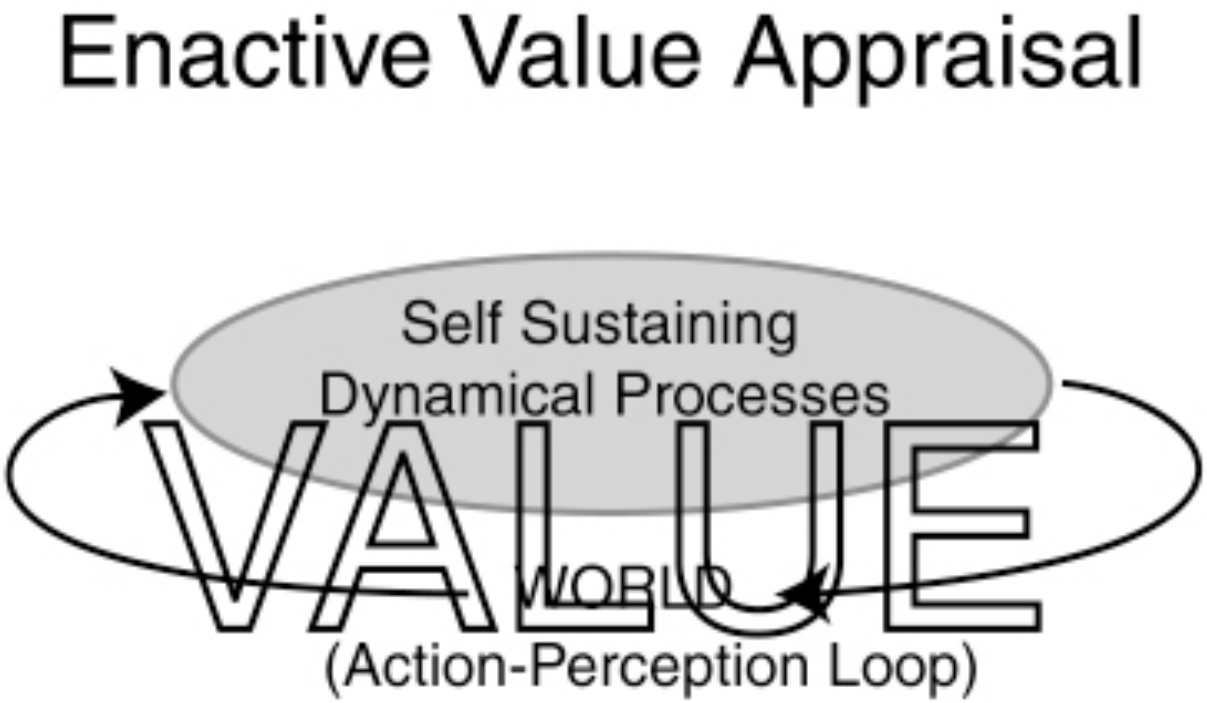
- Dreyfus is known for his extensive critique of GOFAI, but he also takes issue with the field of embodied AI.

- For him the "big remaining problem" is how to incorporate a mechanism of how we "directly pick up **significance** and improve our sensitivity to **relevance**".

- He concludes that such AI models "**haven't a chance** of being realized in the real world".

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology, 20*(2), 247-268

Normativity v3.0: Enactive AI



Di Paolo, E. A. et al. (2010). Horizons for the enactive mind: Values, social interaction, and play. In J. Stewart et al. (Eds.), *Enaction* (pp. 33-87). Cambridge, MA: MIT Press

Natural purpose

- “In such a product of nature every part, as existing through all the other parts, is also thought as existing **for the sake of** the others and that of the whole, i.e. as a tool (organ);

- [. . .] an organ bringing forth the other parts (**and hence everyone bringing forth another**) […];

- and only then and because of this such a product as an *organized* and *self-organizing* being can be called a ***natural purpose***.” (§65)

Kant, I. ([1790] 1987). *Critique of Judgment*. Indianapolis, IN: Hacket Publishing Company

Needful freedom

- "Only living things have needs and **act** on needs.
- Need is based both on the necessity for the continuous self-renewal of the organism by the metabolic process, and on the organism's elemental urge thus **precariously** to continue itself." …
  - "A feedback mechanism may be going, or may be at rest: in either state the machine exists.
- The organism **has** to keep going, because to be going is its very existence – which is revocable – and, threatened with extinction, it is **concerned** in existing." (p. 126)

Jonas, H. ([1966] 2001). *The Phenomenon of Life: Toward a Philosophical Biology*. Evanston, IL: Northwestern University Press.

The hard problem of enactive AI

- Building on Kant and Jonas, the enactive approach to AI includes two necessary requirements:

Systemic requirement	Entailment	*Normativity*
constitutive autonomy	intrinsic teleology	uniform
adaptivity	sense-making	graded

- We called the challenge of realizing these two requirements in an artificial system the "**hard problem of enactive AI**" because it required engineering second-order emergence.

Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence, 173*(3-4), 366-500

The hard problem of enactive AI solved?

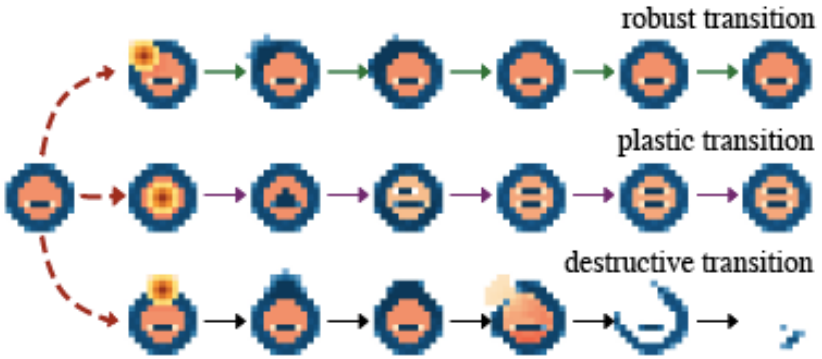


Figure 2: Three perturbations from environment $\mathcal{E}$ applied to *SC*. The perturbations increase the membrane concentration with the same amplitude $\alpha = 2$, as shown in yellow. Following perturbation, the system undergoes transients that stabilize in different attractor classes. The top branch shows a robust transition that returns to the original configuration, the middle branch shows a plastic transition that brings the system to a different stable configuration, and the bottom branch shows a destructive transition.

Agmon, E., Gates, A. J., & Beer, R. D. (2015). Ontogeny and adaptivity in a model protocell. In P. Andrews et al. (Eds.), *Proceedings of the European Conference on Artificial Life* (pp. 216-223). Cambridge, MA: MIT Press

Activity or mere movement?

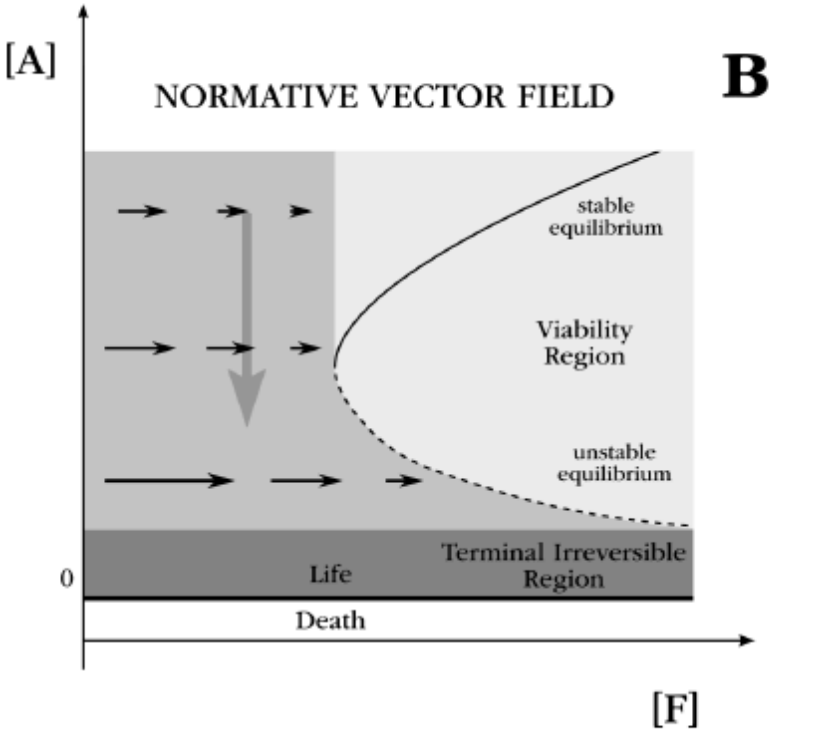
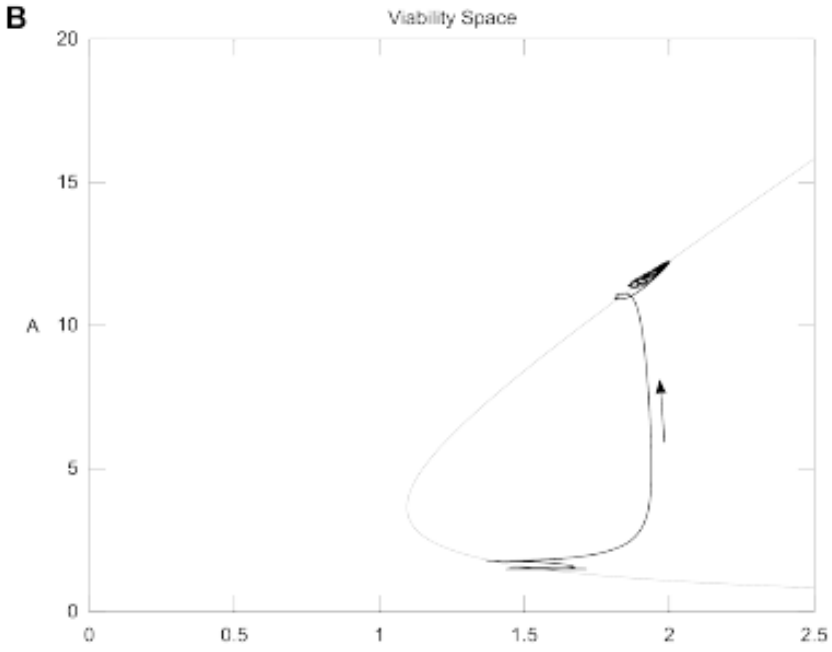


Figure 7: The idea of the 'normative field' in the precarious region – the effects of behaviour as efforts to move the system into the region of viability.



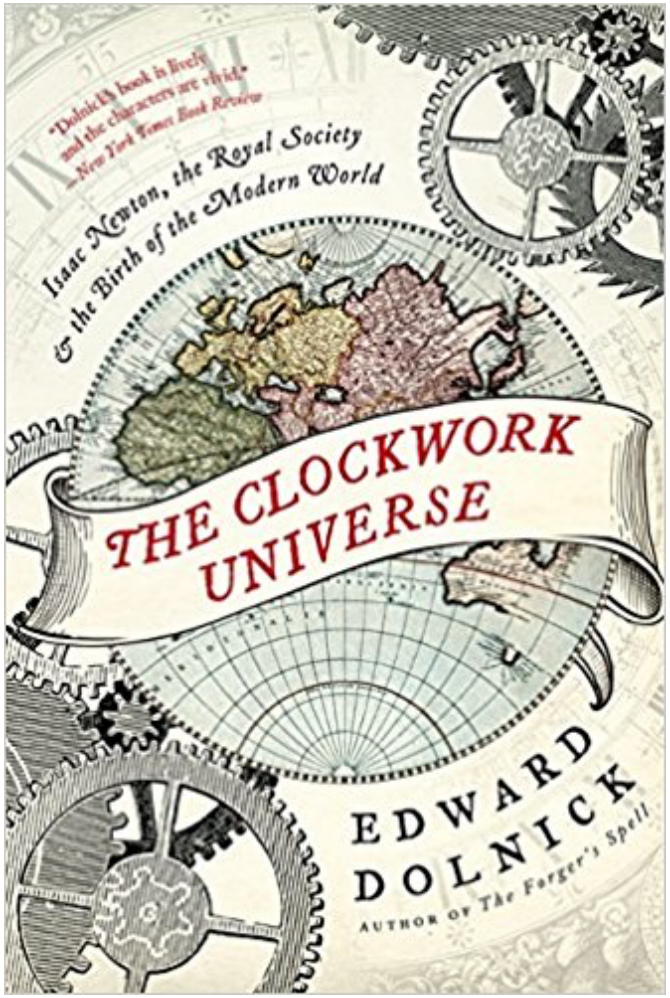
Egbert, M. D., & Barandiaran, X. (2011). Quantifying normative behaviour and precariousness in adaptive agency. In T. Lenaerts et al. (Eds.), *Advances in Artificial Life* (pp. 210-217). Cambridge, MA: MIT Press.

The failure of artificial life

- "Perhaps we have all **missed** some organizing principle of biological systems, or some general truth about them.
- Perhaps there is a way of looking at biological systems which will illuminate an **inherent necessity** in some aspect of the interactions of their parts that is **completely missing** from our artificial systems. …
- I am suggesting that perhaps at this point we simply do not ***get it***." (p. 304)

R.A. Brooks, From earwigs to humans, *Robotics and Autonomous Systems* **20** (2–4) (1997) 291–304.

Science with or vs. Society?

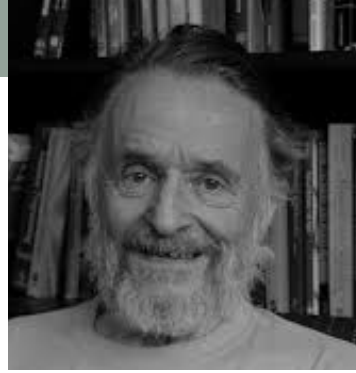


Incomplete nature

- “Each of these sorts of phenomena – a function, reference, purpose, or value – is in some way **incomplete**.
- There is something not-there there. Without this “something” **missing**, they would just be plain and simple physical objects or events […]
- This paradoxical intrinsic quality of existing with respect to something missing […] is irrelevant when it comes to inanimate things, but ***it is a defining property of life and mind***.” (pp. 2-3)
- “It seems that we must explain the **uncaused** appearance of phenomena whose causal powers derive from something **nonexistent**!” (p. 39)

Deacon, T. W. (2012). *Incomplete Nature: How Mind Emerged from Matter*. New York: W. W. Norton & Company

Non-determined nature



- “our theorem asserts that if experimenters have a certain **freedom**, then particles have exactly the same kind of freedom.
- Indeed, it is natural to suppose that this latter freedom is the ultimate explanation of our own.” […]
  - “adding **randomness** also does not explain the quantum mechanical effects” […]
- “The import of the free will theorem is that it is not only current quantum theory, but **the world itself that is non-deterministic**, so that no future theory can return us to a clockwork universe.” (p. 230)

Conway, J. H., & Kochen, S. (2009). The strong free will theorem. *Notices of the American Mathematical Society, 56*(2), 226-232

From wholes to holes?

Bio-machine hybrid AI?

SlugBut (2001)

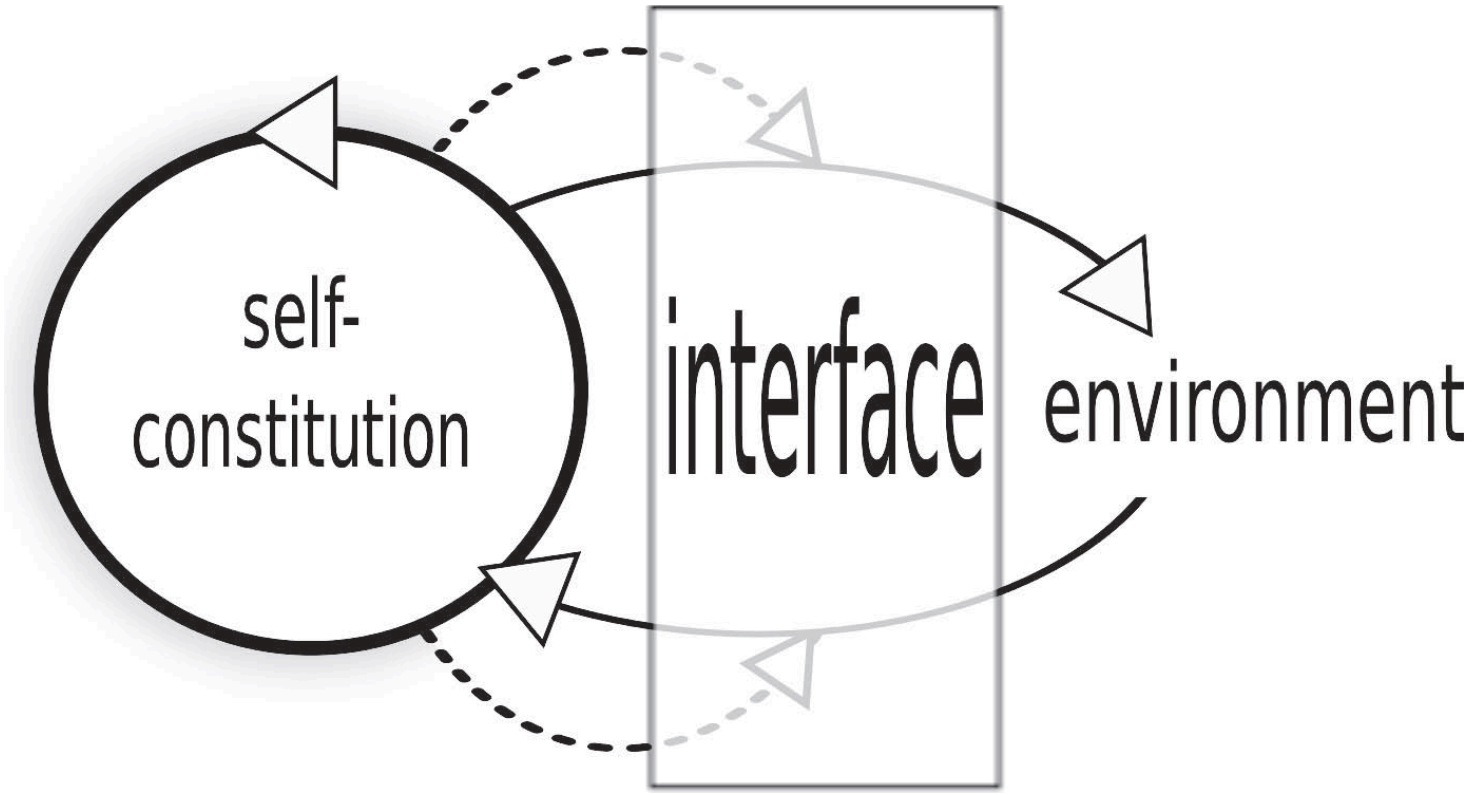
Froese (2014)

Figure 7: EcoWorld finished with EcoBot-III on its robotic track inside. The external (arena) microcontroller is shown on the top, with water and liquid feedstock bottles shown on the left and right, respectively.

EcoBot III (2010)

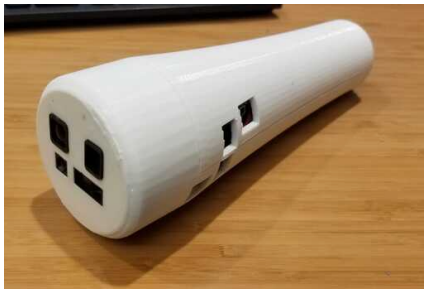
Melhuish et al.

From AI to HCI

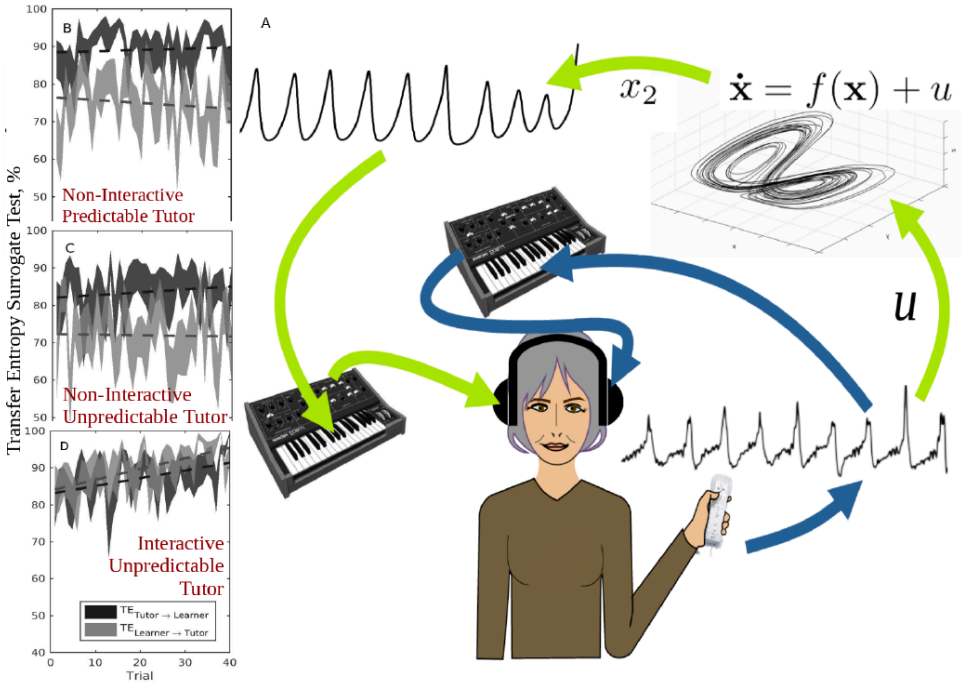


Froese et al. (2012)

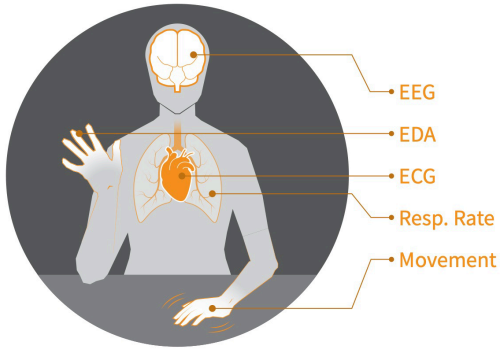
Example projects



Enactive Torch
Froese et al. (2012)



"UNpredictable Interactive system with SONified movement" (UNISON)
Dotov and Froese (2020)



Perceptual Crossing Paradigm
Froese et al. (2020)

More information

- Froese, T. and Ziemke, T. (2009). Enactive artificial intelligence: Investigating the systemic organization of life and mind. *Artificial Intelligence*, 173(3-4), 366-500

- Froese, T. (2014). Bio-machine hybrid technology: A theoretical assessment and some suggestions for improved future design. *Philosophy & Technology, 27*(4), 539-560

- Froese, T. and Taguchi, S. (2019). The problem of meaning in AI and robotics: Still with us after all these years. *Philosophies,* 4(2): 14. doi: 10.3390/philosophies4020014

- Dotov, D., & Froese, T. (2020). Dynamic interactive artificial intelligence: Sketches for a future AI based on human-machine interaction. In J. Bongard et al. (Eds.), *Alife 2020: The 2020 Conference on Artificial Life* (pp. 139-145). Cambridge, MA: MIT Press