The problem of relevance (the frame problem)

PHIL351

Aaron Henry

readings

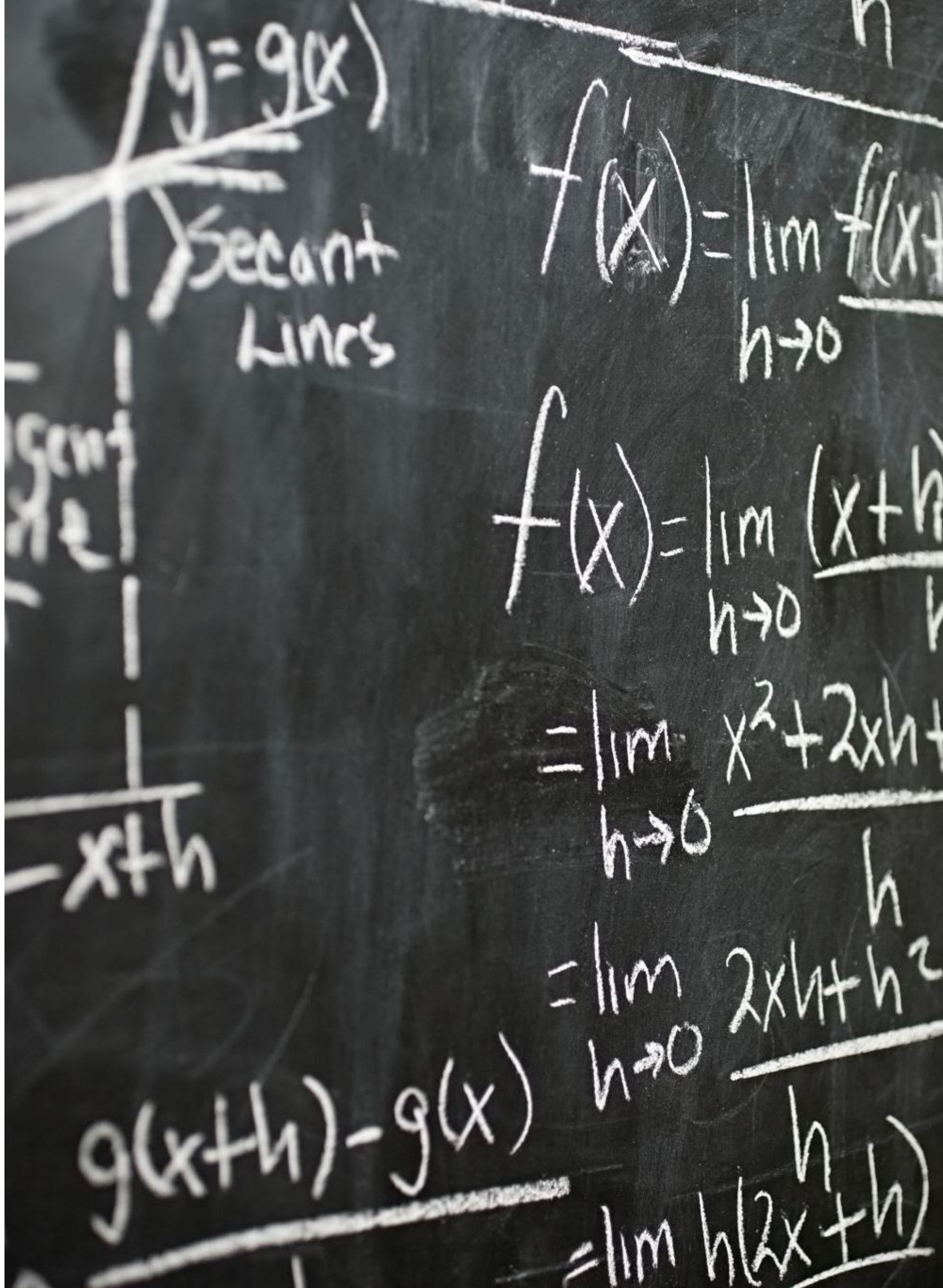
Dietrich et al., Ch. 6 (pp. 137-169)

Optional:

Larson, The Myth of Artificial Intelligence (Chapters 9, 12)

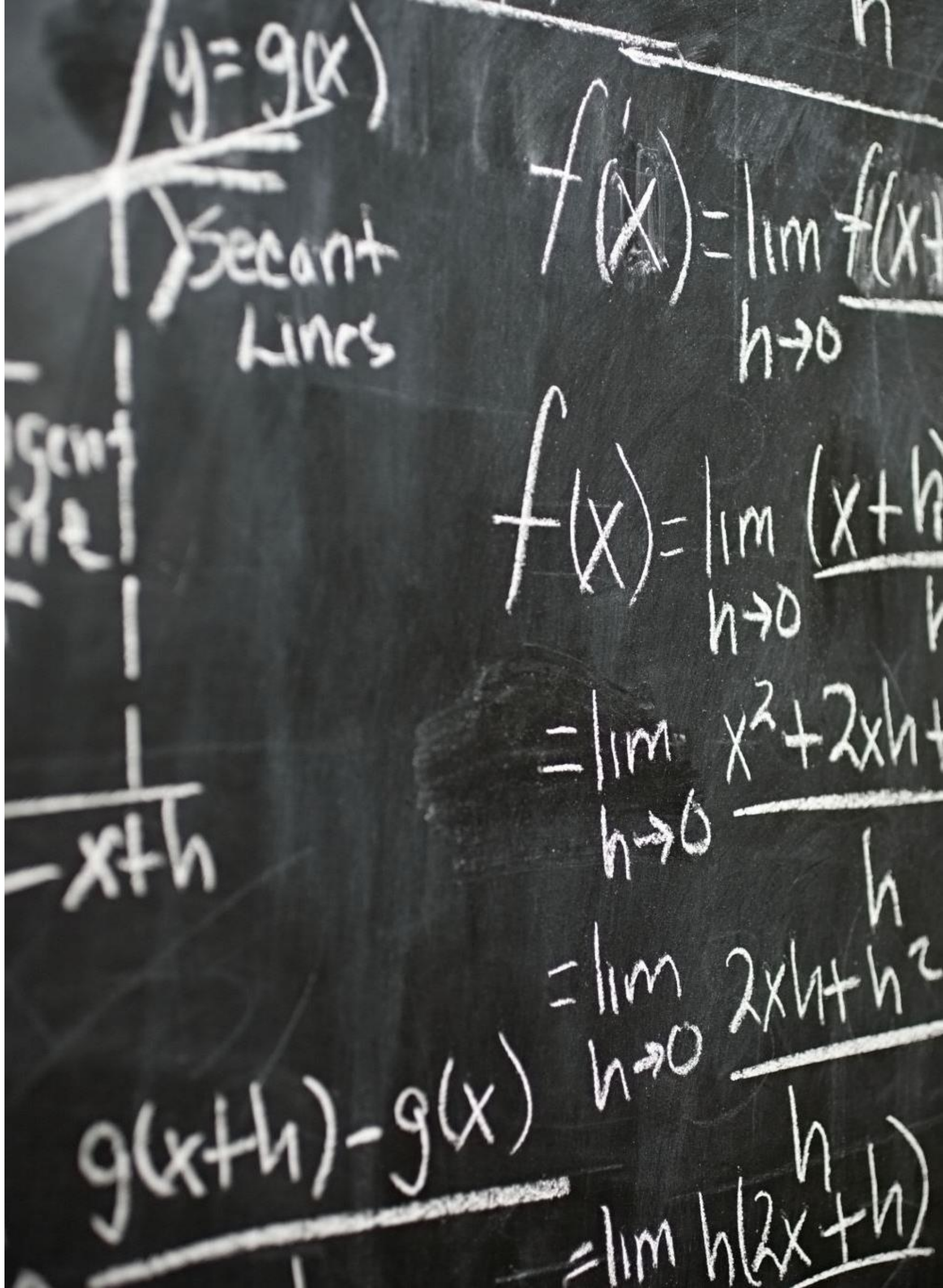
‘The frame problem’

- The phrase ‘frame problem’ is contentious and confusing (see Shanahan, 2006).
- Originally, referred to a technical problem in GOFAI about how to model non-monotonic reasoning (i.e., reasoning that is defeasible in light of changing knowledge) without demanding that the system continually update *every* one of its beliefs. The trick was designing the system so that it updates only beliefs that are relevant.
- Later, philosophers started to use the phrase to describe the problem of explaining ‘common-sense’ and sensitivity to relevance.
- Dietrich et al (2021): computer scientists are just reluctant to admit they have a philosophical problem on their hands (see §3-6)!



‘The frame problem’

- Whereas the problems of meaning and consciousness purport to explain why AI could never succeed in creating genuinely intelligent systems, the frame problem/relevance problem purports to explain why AI research has been unsuccessful to date in creating truly intelligent systems (though some are pessimistic about the prospects of improvement).
- It is more focused on what one can and cannot program a computer *to do* (cf. Dreyfus’s *What Computers Can’t Do*).



The robot parable: Act I

"Once upon a time there was a robot, named R1 by its creators. Its only task was to fend for itself. One day its designers arranged for it to learn that its spare battery, its precious energy supply, was locked in a room with a time bomb set to go off soon. R1 located the room, and the key to the door, and formulated a plan to rescue its battery. There was a wagon in the room, and the battery was on the wagon, and R1 hypothesized that a certain action which it called PULLOUT (Wagon, Room, t) would result in the battery being removed from the room. Straightaway it acted, and did succeed in getting the battery out of the room before the bomb went off. Unfortunately, however, the bomb was also on the wagon. R1 knew that the bomb was on the wagon in the room, but didn't realize that pulling the wagon would bring the bomb out along with the battery. Poor R1 had missed that obvious implication of its planned act …

The robot parable: Act II

“Back to the drawing board. ‘The solution is obvious,’ said the designers. ‘Our next robot must be made to recognize not just the intended implications of its acts, but also the implications about their side-effects, by deducing these implications from the descriptions it uses in formulating its plans.’ They called their next model, the robot-deducer, R1D1. They placed R1D1 in much the same predicament that R1 had succumbed to, and as it too hit upon the idea of PULLOUT (Wagon, Room, t) it began, as designed, to consider the implications of such a course of action. It had just finished deducing that pulling the wagon out of the room would not change the colour of the room’s walls, and was embarking on a proof of the further implication that pulling the wagon out would cause its wheels to turn more revolutions than there were wheels on the wagon—when the bomb exploded …

The robot parable: Act III

"Back to the drawing board. 'We must teach it the difference between relevant implications and irrelevant implications,' said the designers, 'and teach it to ignore the irrelevant ones.' So they developed a method of tagging implications as either relevant or irrelevant to the project at hand, and installed the method in their next model, the robot-relevant-deducer, or R2D1 for short. When they subjected R2D1 to the test that had so unequivocally selected its ancestors for extinction, they were surprised to see it sitting, Hamlet-like, outside the room containing the ticking bomb, the native hue of its resolution sicklied o'er with the pale cast of thought, as Shakespeare (and more recently Fodor) has aptly put it. 'Do something!' they yelled at it. 'I am,' it retorted. 'I'm busily ignoring some thousands of implications I have determined to be irrelevant. Just as soon as I find an irrelevant implication, I put it on the list of those I must ignore, and...' the bomb went off." (Dennett, p. 433).

R1D1, R1D2, and R2D1 seem to lack anything resembling the capacity to *ignore* most of what it knows (/remembers) and selectively attend to only what's currently relevant.

Other morals in the neighbourhood of the frame problem:

- The search space for most real-world problems is vast. To avoid 'combinatorial explosion,' the search for a solution must be *selective*, meaning that one must select not only a solution but a search strategy.
- 'Heuristics and biases': techniques for making search more efficient and computationally tractable at the expense of fallibility ("no free lunch").

Moral of Act 1:

- the robot needs to update its beliefs, both about the intended effects of its actions *and* about side-effects of its actions (e.g., that pulling the wagon will cause the bomb in the wagon to move with it).

Moral of Act 2:

- Telling the robot to update its beliefs about all potential side-effects results in combinatorial explosion.

Moral of Act 3:

- Telling the robot to only update its beliefs about relevant side-effects also results in combinatorial explosion.

In fact, Dennett’s parable understates the problem by only talking about the effects of an action. But there are also all the *non-effects*:

“But here at last is the frame problem. With axioms [like the one we are adding about changing rooms], it is possible to infer what the immediate consequences of actions are. But what about the immediate *non-consequences*? When I go through a door, my position changes. But the color of my hair, and the positions of the cars in the streets, and the place my granny is sitting, don’t change. In fact, most of the world carries on in just the same way that it did before . . . But since many of these things CAN change … we cannot directly infer, as a matter of logic, that they are true [after I change rooms] just because they were [before I changed rooms]: This needs to be stated somehow in our axioms …

In this ontology, whenever something MIGHT change from one moment to another, we have to find some way of stating that it DOESN’T change whenever ANYTHING changes. And this seems silly, since almost all changes in fact change very little of the world. One feels that there should be some economical and principled way of succinctly saying what changes an action makes, without having to explicitly list all the things it doesn’t change as well; yet there doesn’t seem to be another way to do it. That is the frame problem.” (Hayes, quoted in Dietrich et al., p. 147)

Question: Hayes seems to assume that non-effects of an action could be safely ignored, if only one could tell in advance that that is what they are. But is this true? Might information about non-effects sometimes also be highly relevant?

Historical background: the role of abduction in scientific reasoning

- Charles Peirce: ‘abduction’
  - Contrast with deduction and induction (narrowly understood).
- Reasoning (whether practical or theoretical, scientific or everyday, etc.) is a temporally extended process of inquiry motivated by a practical aim.

Historical background: the role of abduction in scientific reasoning

- Some of the stages that Peirce noted:
  - Notice an anomaly/violation of expectation (often, an obstacle to action);
  - Pose a question, framed in relation to our one's interests;
  - Make an inference: a hypothesis/explanation
    - This inference is a 'guess' which leaps to its conclusion but that feels right instinctively;
    - This inference creates a meaningful context for interpreting one's evidence: e.g., the effect becomes a sign of the postulated cause, etc.;
  - Experiment in a 'trial and error' manner
    - making predictions, testing them, revising and updating one's hypothesis to account for prediction errors, drawing connections with other knowledge domains and particular events, using the analogy to guide the collection of new data, etc.);
  - If all goes well, one arrives at a workable (but still defeasible) explanation of the phenomenon that facilitates more fluent interaction with the world.

Historical background: the role of abduction in scientific reasoning

- Karl Popper (following Peirce):
  - The scientific method is a process of conjecture and refutation ('falsification');
  - The conjecture is a creative leap of the imagination (unsupported guess-work), followed by rigorous empirical investigation (logically derive a bold prediction, then test it via controlled experimentation);
  - 'Fallibilism': empirical theories can never be conclusively confirmed but only disconfirmed. Science progresses through the incremental elimination of falsity.

Historical background: the role of abduction in scientific reasoning

- But as later philosophers (e.g., Quine, Goodman) noted, Popper’s two steps cannot be so cleanly separated:
    - There is no perfect experiment to test an isolated hypothesis. We bring to bear our entire ‘web of belief’ in interpreting some observation.
    - Depending on how integrated a belief is within one’s belief system, one might treat the false prediction as evidence against the belief in question or as evidence that one’s measurement instruments are imperfect, need recalibration, etc.
- Moral: theory confirmation is holistic, open-ended (endlessly revisable in light of future observation), and apparently un-codifiable.

Historical background: the role of abduction in scientific reasoning

- These points seem to apply to our everyday (non-scientific) thinking and reasoning too, much of which we do automatically and unconsciously.
    - The ease and speed of an inference is not an indication of its triviality!
- A special case of everyday abductive reasoning: interpreting another agent’s intentions, including communicative intentions. To successfully interpret what a person is saying or doing, we must be attuned to the same background of relevance as they are (cf. ‘pragmatics’).

Fodor: the frame problem and scientific rationality

- Jerry Fodor has persistently stressed the difficulty of the frame problem, typically by reference to salient features of scientific rationality.
- On his view, while *modules* are plausibly computational, *central systems* are almost certainly not.
  - Modules ≈ 'mono-intelligences' (Dietrich et al.) and what many today have in mind when discussing 'Narrow AI.'
  - Central systems ≈ 'open-ended intelligence' (Dietrich et al.) or 'general intelligence'.

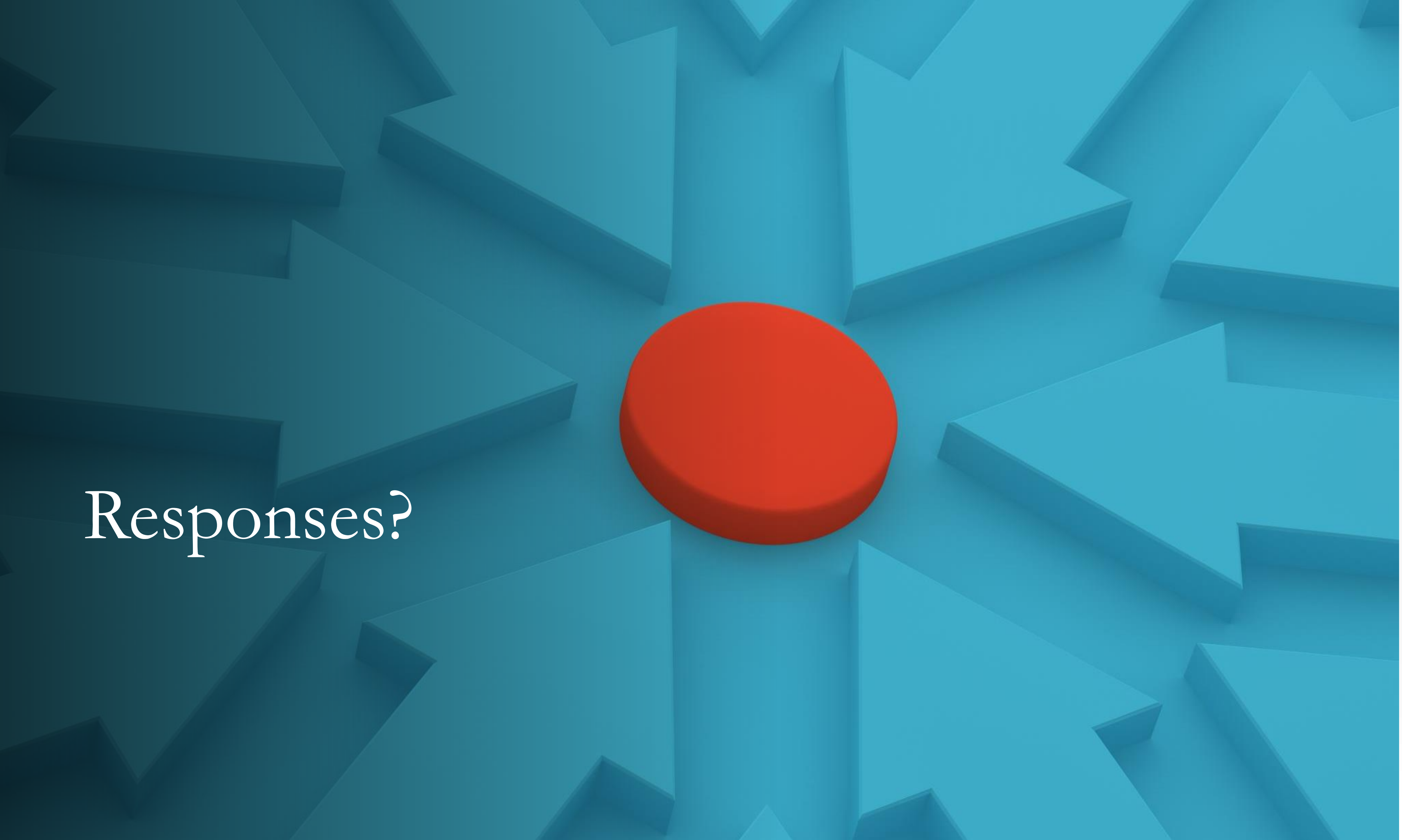
Fodor: the frame problem and scientific rationality

Fodor highlights two (intimately related) properties of central cognitive states that resist computational explanation:

- Being 'Quinean': epistemic/rational properties of an attitude that are relative to a larger set of attitudes. (Fodor 1983, 107)
  - E.g., coherence, simplicity, plausibility are Quinean because they are relative to an entire belief system.
  - Quineanism generates 'the globality problem' (Schneider 2011): computational operations are sensitive to a mental state's local (syntactic) properties but not to its global (e.g., Quinean) properties.
- Being Isotropic: any member of an attitude set is potentially relevant to any other. (Fodor 1983, 105)
  - Analogical reasoning (e.g., thinking of the structure of the atom as like that of a solar system, and of evolution as like artificial selection and like a market economy)
  - Isotropism underlies 'the relevance problem' (Schneider, 2011): computational operations are insensitive to relevance.

“I want to suggest two morals before I leave this point . The first is that the closer we get to what we are pretheoretically inclined to think of as the ‘higher,’ ‘more intelligent’, less reflexive, less routine exercises of cognitive capacities, the more such global properties as isotropy tend to show up. I doubt that this is an accident. I suspect that it is precisely its possession of such global properties that we have in mind when we think of a cognitive process as paradigmatically intelligent. The second moral preshadows a point that I shall jump up and down about further on. It is striking that, while everybody thinks that analogical reasoning is an important ingredient in all sorts of cognitive achievements that we prize, nobody knows anything about how it works; not even in the dim, in-a-glass-darkly sort of way in which there are some ideas about how confirmation works. I don't think that this is an accident either. In fact, I should like to propose a generalization; one which I fondly hope will some day come to be known as ‘Fodor's First Law of the Nonexistence of Cognitive Science’. It goes like this: the more global (e.g., the more isotropic) a cognitive process is, the less anybody understands it. *Very* global processes, like analogical reasoning, aren't understood at all.” (Fodor 1983, p. 107)

This from the same person who has claimed, just as persistently, that the computational theory of mind is “the best theory we’ve got” (Fodor, 2000)!



Responses?

A change of paradigm?

Paradigm shift #1: ANNs

- ANNs might seem a more promising an architecture than classical GOFAI for addressing the frame problem:
    - Greater ease dealing with context and global properties of stimuli;
    - Memory is an evolving state of the network implicitly in the connection weights.
- However, there are many reasons to doubt that a move to ANNs is sufficient to solve the frame problem, especially the relevance problem.
- As statistical engines, ANNs tend to assimilate novel cases to pre-existing categories, resulting in 'overfitting' to the patterns present in its training data.
    - Cf. adversarial examples

A moral: a plea for Socratic ignorance?

- We want a system that will not only recognize anomalies as such but that will recognize certain anomalies as warranting inquiry, where inquiry involves exploiting a model of the world to solve a problem while remaining cognizant of its status *as* a model (and so always provisional and potentially revisable).
- Witness the role of ignorance in abductive reasoning:
  - To reason analogically, the system must see the target domain as something about which it is ignorant but which it might better understand through creatively repurposing its existing concepts.
  - To reason causally, the system must see an effect as a particular with an idiosyncratic history about which it is (and will remain) ignorant but which it can better understand by entertaining plausible hypothetical scenarios that are consistent with one's model of the world and that can isolate causally relevant variables of a situation.

Paradigm shift #2: 4E cognitive science

- Recall that the frame problem is, in part, the problem of building a system that will behave intelligently in real time.
- 4E (embodied, embedded, extended, and enactive) cognitive scientists start from the premise that the intrinsic temporal dynamics of cognition is one of its most important properties. This provides the impetus to consider the contributions of a system's embodiment and environment to the performance of cognitive tasks. The more cognitive labour can be 'off-loaded' to the agent's body and environment, the less reason we have to fear the frame problem!
- In its more radical forms (e.g., enactivism and ecological psychology), 4E cognitive scientists reject the idea that cognition is properly understood as computation partly because this misinterprets the temporal structure of cognition.

Situated robotics

- Rodney Brooks' 'subsumption architecture' robotics put forward as an alternative to the (then dominant) architecture, relying on iterated sequences of 'sense-model-plan-move' (e.g., SHAKEY).

- In Brooks' robots, the sensors were directly wired to various 'activity layers' (e.g., WANDER, AVOID, EXPLORE), which would compete via mutual inhibition. When an activity layer wins out, the sensors directly and continuously drives behaviour proprietary to that layer (with no mediating model and plan). Whereas SHAKEY's behaviour was ploddingly slow and brittle, Brooks' robots gracefully navigated cluttered environments, often in surprisingly flexible, overtly goal-directed ways.

- Brooks claimed his robots did not have any use for representations. ("the world is its own best model").

From existential phenomenology to … the Roomba.

Thinking through coupling

Mind/cognition spans the (tightly coupled) agent-environment system:

> “It is only for convenience (and from habit) that we think of the organism and environment as separate; in fact, they are best thought of as comprising just one system” (Chemero 2001, p.142).

> “The emergence of mind takes place in the medium of patterns of activation across neuronal cell assemblies in conjunction with the interaction of their attached sensors (eyes, ears, etc.) and effectors (hands, speech apparatus, etc.) with the environment in which they are embedded. Make no mistake about it, *that* is the stuff of which human minds are made: brains, bodies, and environments.” (Spivey 2007, 33, emphasis original).

Dynamical systems theory

A branch of mathematics used to model system whose behaviour changes continuously over time, including patterns of stability, instability, and meta-stability. Uses concepts like ‘attractor’, ‘phase shift’, ‘coupled variables’, and ‘self-organizing criticality’ to model the dynamical system’s behaviour.

- An alternative set of formal tools for describing the activity artificial neural networks and autonomous robotics.
- Applications to human and animal cognition:
    - the outfielder problem;
    - the diving gannet (which must dive from 100 ft aiming for its prey below the water) whose sensorimotor dynamics are well-described in terms of ‘τ’ (the ratio of the size of a projected image to the rate of change of the image’s size, which gives information about time-to-contact, see Chemero 2009, pp. 123-4).
    - modelling cognitive development in infancy (e.g., locomotion; A-not B error) (Thelen & Smith 1994).

Enactive and ecological cognitive science

- According to the ‘enactive’ approach to cognitive, both meaning and relevance are ‘enacted’ or brought forth through an organism’s active self-maintenance (autopoiesis). Similarly, ecological psychologists insist that the information to which organisms are perceptually attuned is ‘ecological information,’ which is relative to an organism’s morphology.

- Common theme: properties like meaning and relevance do not belong to internal states of an agent (e.g., representations of mind-independent objects and properties) but to an organism’s *milieu*: agent and environment are complementary aspects of an integrated system.

- If so, cognitive agents do not need to solve the frame problem since, in virtue of being alive, they already inhabit a world of salience and significance: elements of the environment stand out as ‘affordances’ with an inherent biological significance to the organism.

Critiques of enactive and ecological approaches

- Enactive and ecological theorists seem to face a frame problem of their own, given the huge number of affordances available to an organism at any time.
    - How does the organism become attuned to only the relevant affordances?
- Is biological agency sufficient for cognitive agency?
    - Intelligence comes into play when what matters to an organism is something to which it is not currently causally coupled ('representation-hungry' tasks)
    - Cognitive agents do not engage with the world *merely* as an arena for action but also as an objective world to which one defers (cf. Smith).

The massive modularity hypothesis

Humans only give the appearance of engaging in domain-general cognition. We can explain the manifest flexibility of human cognition without having to postulate any unencapsulated computational processes (of the sort that would land us in the frame problem).

- Mind as a ‘Swiss-army knife’ or a ‘bag of tricks’ – i.e., a set of appropriately organized special-purpose devices. (e.g., Cosmides & Tooby 1992; Pinker 1997; Carruthers 2006; Sperber 2012).
- No *truly* domain-general cognition (at least not in the way that Fodor supposed). The observed flexibility of human cognition is to be explained in other terms (or explained away).
    - Connection with ‘frames,’ ‘scripts,’ ‘schemas,’ and to nativism.
- Mind as a ‘Swiss-army knife’ or a ‘bag of tricks’ (i.e., a set of appropriately orchestrated special-purpose machines)

The massive modularity hypothesis

Proponents of MM (e.g., Cosmides & Tooby 1992) relied heavily from the heuristics and biases literature to argue that human reasoning is more content-specific than we traditionally assume (e.g., Wason selection task).

2	7	E	Z

Vodka	Water	50 years old	6 years old

X is drinking alcohol	X is 19 years or older	X is drinking alcohol → X is 19 years or older
T (e.g., X is drinking beer)	T (e.g., X is 45 years old)	T
T (e.g., X is drinking vodka shots)	F (e.g., X is 10 years old)	F
F (e.g., X is drinking water)	T (e.g., X is 75 years old)	T
F (e.g., X is drinking water)	F (e.g., X is 12 years old)	T

P	Q	$P \rightarrow Q$
T	T	T
T	F	F
F	T	T
F	F	T

The massive modularity hypothesis

- In his *The Mind Doesn't Work That Way* (2000), Fodor argues (among other things) that the frame problem re-arises for these accounts: when a situation has features that are consistent with many different task domains, how is the relevant module selected? That is, how does the system determine what context it is in?
  - Cf. Dreyfus' *What Computers Can't Do*

Solving the frame problem via consciousness?

- A final option is to uphold the traditional view of human cognitive architecture as divided into peripheral and central systems (lower and higher, automatic and controlled, etc.), and claim that we will explain the distinctive flexibility of central systems via a cognitive theory of *consciousness*.
- Two dominant cognitive theories of consciousness:
    - Global workspace theory
    - Higher order theories

Global workspace theory of consciousness

Against MM, proponents of GWT accept that humans engage in genuinely domain-general cognition.

Further, GWT sees domain-general cognition as explanatorily related to consciousness (e.g., Baars 1988; Shanahan & Baars 2005; Shanahan 2007; Schneider 2011).

(Dehaene & Naccache (2001) later renamed the theory 'Global Neuronal Workspace Theory' to reflect increased integration with neuroscience.)

Avoiding combinatorial explosion

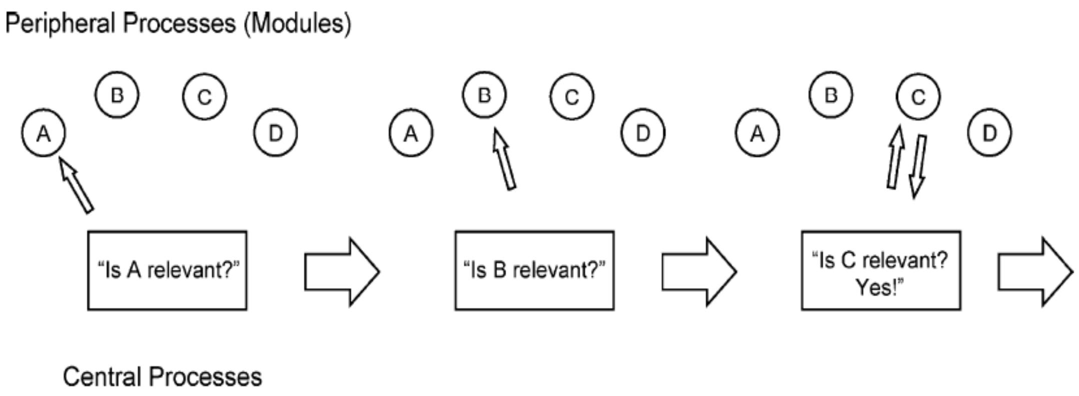


Fig. 2. A naïve model of information flow.

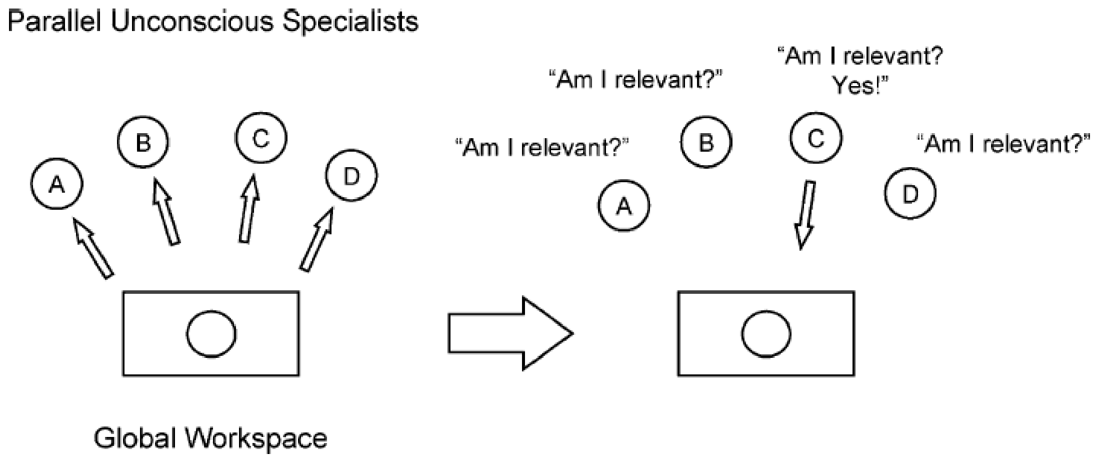
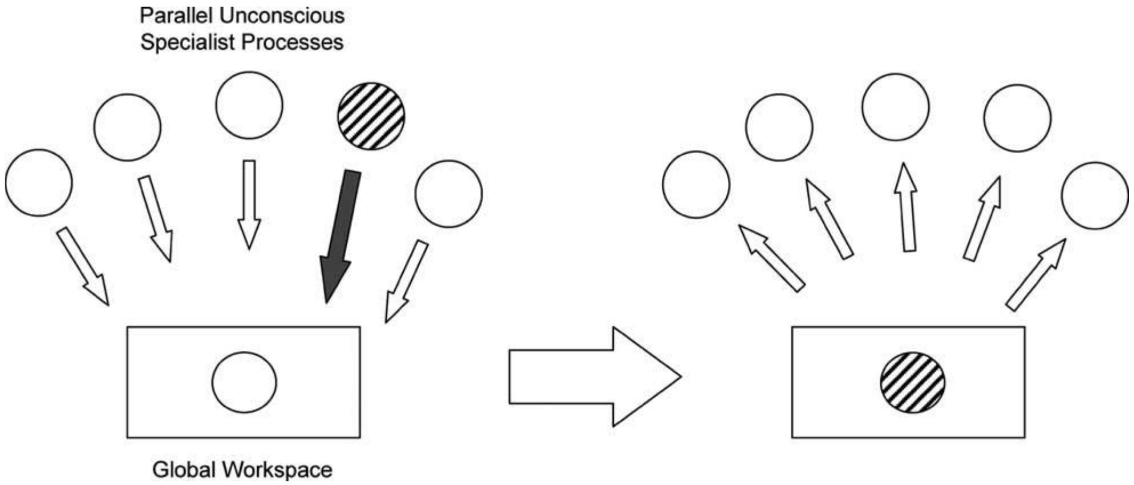


Fig. 3. The global workspace model of information flow.

Global workspace theory

Domain-general cognition is computationally feasible because it is serial in only a restricted sense. The contents of consciousness are the outcome of a competition between a massive number of modules operating in parallel, the winner of which has its content globally broadcast to the rest of the cognitive system.



Global workspace theory

GWT claims to solve the frame problem in a manner fully consistent with classical computational cognitive science.

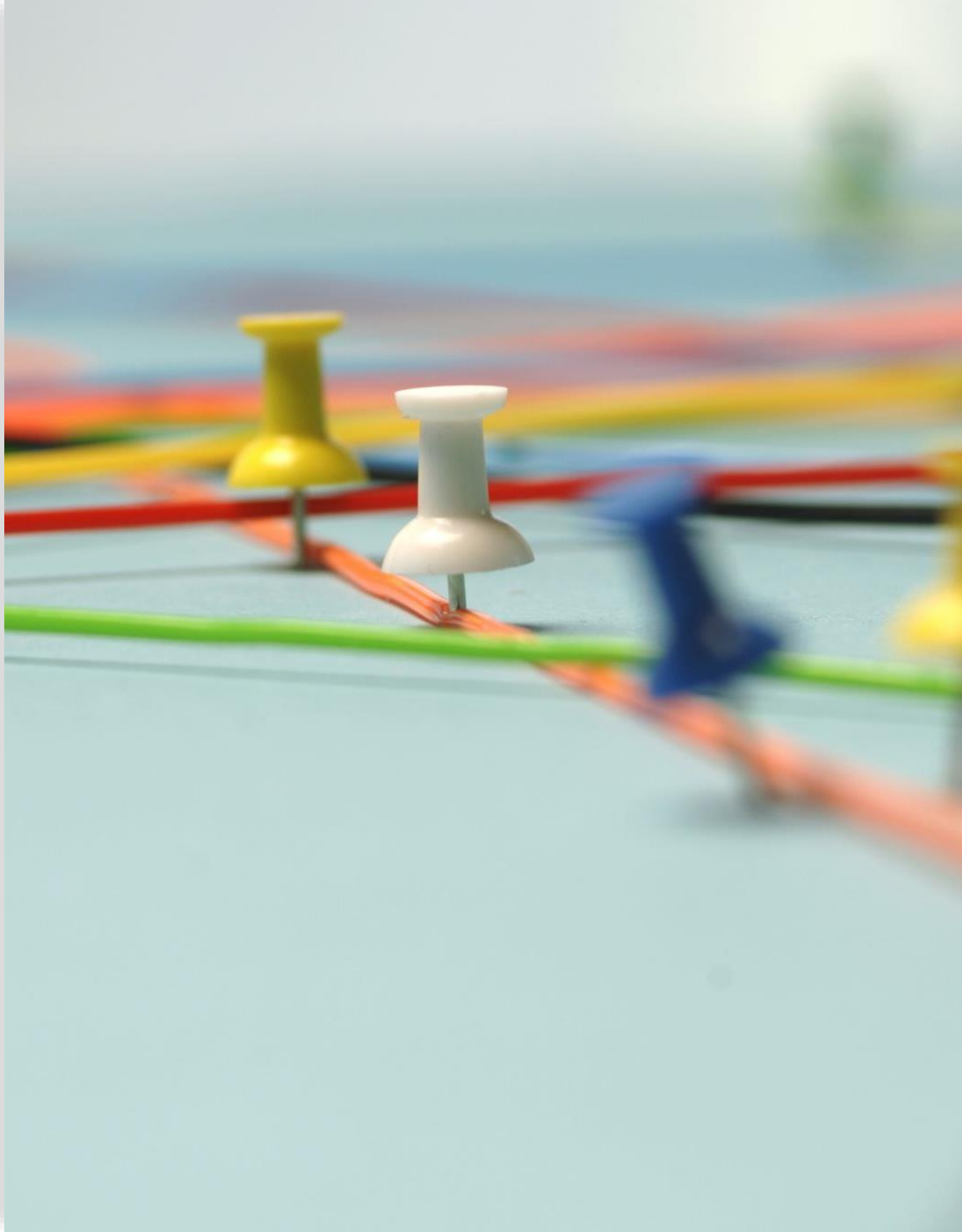
- *Globality* is explained by the integrative coordinative effects of global broadcasting (achievable through a "small-world" neural architecture, see Shanahan 2007).

- *Relevance* is explained by competitive interactions between modules in response to the globally broadcasted content (see Shanahan & Baars 2005)

Global workspace theory

Question: by what process is the competition's 'winner' determined and allowed access to the global workspace?

The standard answer invokes an "attentional mechanism" as the gatekeeper of the global workspace (Shanahan & Baars 2005; Prinz 2012). But, to play such a role, it seems attention must be capable of detecting relevance. How does this work?



Higher order theories of consciousness and self-modelling

- Higher Order theories of consciousness: a conscious state is a mental state *of* which one is conscious, i.e., that one represents oneself as being in.
  - The higher order representation (meta-representation) might be a thought (HOT) or percept (HOP).
  - The higher order representation (meta-representation) might be distinct from the first-order representation (in which case the conscious mental state is the represented mental state) or identical to the first-order representation (in which case the conscious mental state is self-representing).
- A related idea: a conscious state is one which embeds a certain representation of the self within its content (with different proposals offering different proposals about 'the self').
- A common theme: consciousness is not a mere 2-place relation between X (a representation) and Y (a represented object) but a 3-place relation between X, Y, and the subject *to* or *for* whom Y is represented via X, where 'the subject' is itself explained in representational terms.

Potential relevance to the frame problem?

- As Dietrich et al. note, cognitive agents do, often enough, fall victim to the frame problem when they fail to notice what's relevant about the situation they are in. As their examples illustrate, such failures are often due to some limitation in the cognitive perspective that a subject (or community of subjects) has adopted on their situation, locking them into a certain cognitive framing that restricts their problem-solving capacity.
    - Cf. functional fixedness
- Embedding a representation of oneself and/or one's current cognitive perspective on a problem *within* one's overall representation of the problem might afford the agent with opportunities to reflectively 'step back' from their cognitive framing of a situation in order to reflectively assess its appropriateness to the task at hand and adjust their perspective in creative ways (cognitive restructuring), enabling greater cognitive flexibility.
    - (probably not a sufficient condition)

Option to integrate GWS and HO?

- Some have attempted to reap the benefits of both GWS and HO approaches by theoretical synthesis (e.g., Van Gulick). On such a view, the representational contents that gain entry to the global workspace become meta-representational in the process, such that the state becomes functionally integrated with the rest of the cognitive system via global broadcasting and a site of metacognitive agency.
  - Global broadcasting might capture the unity of consciousness;
  - Higher order representation might capture the subjectivity, or 'for-me-ness,' of consciousness;

From the frame problem to the problem of consciousness …

- While it is promising to think that one or another of these cognitive theories of consciousness might offer a (partial or complete) solution to the frame problem, they also land us in another philosophical minefield – namely, the prospects of a science of consciousness.
- It is to that philosophical minefield that we now turn …