Phenomenology and the Cognitive Sciences
https://doi.org/10.1007/s11097-024-09979-6

Framing the predictive mind: why we should think again about Dreyfus

Jack Reynolds[1]

Accepted: 25 March 2024

Abstract

In this paper I return to Hubert Dreyfus' old but influential critique of artificial intelligence, redirecting it towards contemporary predictive processing models of the mind (PP). I focus on Dreyfus' arguments about the "frame problem" for artificial cognitive systems, and his contrasting account of embodied human skills and expertise. The frame problem presents as a prima facie problem for practical work in AI and robotics, but also for computational views of the mind in general, including for PP. Indeed, some of the issues it presents seem more acute for PP, insofar as it seeks to unify *all* cognition and intelligence, and aims to do so without admitting any cognitive processes or mechanisms outside of the scope of the theory. I contend, however, that there is an unresolved problem for PP concerning whether it can both explain all cognition and intelligent behavior as minimizing prediction error with just the core formal elements of the PP toolbox, *and* also adequately comprehend (or explain away) some of the apparent cognitive differences between biological and prediction-based artificial intelligence, notably in regard to establishing relevance and flexible context-switching, precisely the features of interest to Dreyfus' work on embodied indexicality, habits/skills, and abductive inference. I address several influential philosophical versions of PP, including the work of Jakob Hohwy and Andy Clark, as well as more enactive-oriented interpretations of active inference coming from a broadly Fristonian perspective.

Keywords Hubert Dreyfus · Frame problem · Predictive processing · Cognition · Predictive AI · Embodiment · Skills · Active inference

In this paper I return to Hubert Dreyfus' old but influential critique of artificial intelligence, redirecting it towards contemporary predictive processing models of the mind (PP). I focus on Dreyfus' arguments about the "frame problem" for artificial

✉ Jack Reynolds
Jack.reynolds@deakin.edu.au

[1] Faculty of Arts and Education, Deakin University, 221 Burwood Highway, 3125 Burwood, Australia

Published online: 06 May 2024

cognitive systems, and his contrasting account of embodied human skills and expertise, which he thinks avoid the problem of how to comprehend background meaning, and of establishing and updating determinations of relevance in dynamic real-world scenarios. Although most of the major PP theorists have not explicitly claimed to have solved the frame problem, nor systematically addressed it[1], it is arguable that it ought to be addressed by any view that purports to comprehensively explain mind and cognition via purely predictive and computational means. The frame problem presents as a prima facie problem for practical work in AI and robotics, but also for for computational views of the mind in general. If all cognition is computation, even if probabilistically construed, there is no reason to expect any constitutive barriers to artificial general intelligence, notwithstanding the complexity of the task. This issue also seems to be acute for proponents of PP, since many of them aim to computationally unify *all* cognition and intelligence, and to do so without remainder: that is, without admitting any cognitive processes or mechanisms outside of the scope of the theory. While PP's account of cognition is very different from that criticised by Dreyfus as underpinning "Good Old-Fashioned AI" (GOFAI), the terms of Dreyfus' critique help to highlight an unresolved problem concerning whether PP can both explain all cognition and intelligent behavior as minimizing prediction error with just the core formal elements of the toolbox, *and* also adequately comprehend (or explain away) some of the apparent cognitive differences between biological and prediction-based artificial intelligence in regard to establishing relevance and flexible context-switching (i.e. general intelligence).

This paper proceeds with the following sections:

1. Dreyfus and the frame problem
   First, I outline Dreyfus' views regarding the frame problem for artificial general intelligence, and his contrasting account of human intelligence, learning, and skill-acquisition. Together these present a prima facie obstacle for computationalism, at least pending empirical results in open and dynamic real-world environments;
2. Predictive processing and the frame problem
   I then seek to establish the broad applicability of the frame problem to PP's account of mind and cognition, notwithstanding key PP additions like the idea of "active inference", "hyper-priors", and the fact it offers a probabilistic construal of cognition via Bayesian prediction error minimisation. The argument about the frame problem hinges on the difficulty for PP in computationally accommodating three connected factors that were central to Dreyfus's work: embodied indexicality, habits/skills, and problem-solving abductive inference. I argue that these are not as obviously amenable to a strict Bayesian and active inference construal as proponents of PP claim.
3. Replies and objections

[1] There is, however, important consideration of the frame problem in Froese and Ziemke (2009), Linson et al. (2018, 14–15), Froese and Taguchi (2019), Andersen et al. (2022), and Kiverstein et al. (2022). In general I concur with the spirit of much that work, and address it further in Section 3 below.

Finally, I consider some possible replies and objections to my "framing" of the predictive mind. These concern: the *pluralism* of some versions of PP; the idea that the *embodied organism is itself the model*; and the claim that *relevance just is precision-weighting*. While I agree with the overall direction of this work in embodied PP, I argue that these responses typically concede that PP's predictive and inferential form of computationalism is not wholly sufficient, by itself, and that there is an ambiguity around the formal relationship between abduction and Bayesian inference that has not yet been convincingly addressed.

1 Dreyfus, the frame problem and GOFAI

Hubert Dreyfus is well-known for his critique of GOFAI, even if this critique was much more positively received in some circles than in others. In brief, Dreyfus held that GOFAI models of the mind (symbolic and computational) would be unable to emulate or surpass human intelligence in many dynamic real-world contexts, essentially due to the frame problem. For Dreyfus, the frame problem concerns how any information processing system might quickly and flexibly sort relevant from irrelevant cognitive processes and information, without some pregiven "frame" or script, which was computationally intractable for GOFAI. It would also defeat the purpose of the very idea of AGI, *if* it required a programmer to demarcate frames and rules for their application.[2] Dreyfus' critique left open the possibility that other computational models of cognition (non GOFAI) might resolve these difficulties, however, whether connectionist, Bayesian-based PP, active inference, etc. Other models may avoid the frame problem precisely by virtue of being less symbol-oriented and deductive. Despite this possibility, Dreyfus' body of work still presents as a prima facie obstacle – perhaps to be dissipated or overcome – for views of the mind as computer-like, one that can be re-examined today, more than 50 years since he published *What Computers Can't Do*.

Dreyfus framed the apparent difference between human and artificial cognitive systems as being about the (in)capacity of AI to emulate our context-dependent "common-sense", or even "intuition" (which we might also understand as "creative abduction" in C.S. Peirce's terms), as well as skilled human "expertise". On Dreyfus' portrayal that expertise is intuitive rather than rule-governed, but rather than this being especially mysterious, intuition and skilled judgment is grounded in the context-sensitive and holistic nature of our embodied habits. These embodied habits and skills scaffold common-sense knowledge and flexible problem-solving –but for GOFAI, and arguably also AI today, algorithmically programming such capacities was much more difficult than anticipated, with potential issues regarding new contexts and establishing what stored memory/information needs to be updated and what does not, as well as problems of infinite regress (rules about the application

[2] In its original incarnation, the frame problem concerned how to formalise the non-effects of an action (McCarthy & Hayes, 1969). Dreyfus, Dennett, Fodor and others gave it a wider reach, beyond its original formal reasoning context.

of rules), and of deixis (establishing the 'here' and 'now'). Dreyfus argued that any context-independent symbols aiming to serve as representational states of a cognitive system will not emulate the sorts of contextualized practical understanding that humans have in navigating and cognizing our worlds. An early problem with GOFAI he pointed to was the extreme difficulty of programming an AI system to understand children's stories using the sort of background common-sense that enables a five-year old to comprehend and engage with that material (Dreyfus, 1992, x, 57–62), both in understanding jokes and in understanding potential suspects in a simple 'who done it?' scenario (I will return to this).

While the frame problem was not mentioned by name in the first version of *What Computers Can't Do*, it was implicitly there. It was then emphasised in the book's 1979 revision, as well as in his later book with his brother, Stuart (Dreyfus & Dreyfus, 2000). In these works, Dreyfus drew attention to two basic and inter-connected issues he thought central to the problem: *meaning holism*, and *embodied skills and know-how* (Dreyfus, 1992, xii; cf. Wheeler, 2005, 174). Dreyfus' account of meaning holism is indebted to the philosophy of Heidegger, who argued in *Being and Time* that contexts are complex, network-like semantic structures defined with reference to the concerns and projects of an agent, and which also includes social norms. For Heidegger, the example of choice is the hammer, and the holistic "equipmental nexus" that obtains between a hammer, a nail, a planer, etc., given the project of fixing a fence. And while there is a holistic connection between a series of potential objects that are relevant for a given project like fixing a fence, there are also more marginal connections that form part of the network of association, or the "field" of affordances as Bruineberg & Rietveld put it (2014), borrowing from both J. J. Gibson and Dreyfus. This idea of a field of affordances is available when humans need to context-switch. It is hence flexible, rather than being single-tracked and brittle in the face of changing circumstances. For example, I may be working at home on my fence with a Heideggerian hammer. I may be ensconced in my hammering activity and its field of directly relevant affordances, but the sound of the dog barking for food, or of the washing machine finishing up (and recognition of a need to get the clothes on the line and dry) might solicit my attention. Some of these affordances have a direct environmental trigger, but others are subtler and less obvious, whether for human or animal cognition. Something which presents as an affordance in one context need not be so in another: i.e. not smoking a cigarette in hospital in Gilbert Ryle's example, or the rabbit who sees their burrow as alternatively a place to flee or to sleep depending on the context of predators and the time of the day (cf. Dreyfus, 2007), or a bottle of water that might solicit my attention during my lecture if I placed it there, but may not do so if it was already in the theatre before I began (Bruineberg et al., 2018). Although humans make all sorts of formal reasoning errors, this context-sensitivity is a strength.

For Dreyfus, the relevant associations and understanding of the semantic structures are bound up with embodied skills (1992, xix). The five-year old's background knowledge or "common sense" consists in a set of practical knowledge and skills, which depend on a body and its relation to the world. Our bodies cannot do anything at all, of course, since we are morphologically and kinetically constrained. They open up particular possibilities for us, a motor intentionality in terms of what we

can do, and they preclude others. Our proprioceptive bodies also inaugurate a "here" and a "now", allowing us to readily comprehend relational features, like near and far, within reach and out of reach, up and down, and apprehend that something is larger or smaller than something else. They thereby assist us determining relevance/irrelevance in regard to the frame problem. We do not need to deliberately reason that a predator is a threat because it is proximate, or not a direct threat because it is far away. This kind of embodied know-how pervades biological cognition in vertebrates and invertebrates too (cf. Godfrey Smith, 2020). An animal also comprehends a potential predator as near or far, and understands in an embodied and lived manner that a decent-sized river separating them from a potential predator may render them safe, even if the predator is objectively quite close[3]. As Dreyfus put the point: "If everything is similar to everything else in an indefinitely large number of ways, what constrains the space of possible generalizations so that trial and error learning has a chance of succeeding? Here is where the body comes in" (Dreyfus, 1998).

By contrast, Dreyfus contended that GOFAI could not properly address the problem of deixis (1992, xx), representations that refer to the 'here' and 'now', as well as vaguer references like "over there" or "nearby". This resulted in ongoing difficulties for AI systems in locating objects with respect to their own location in the world. Daniel Dennett dramatically illustrated the importance of this issue for determining what is and is not "relevant" (Dennett, 2006). In Dennett's story, a robot (R1) learns that (a) its energy supply is located in a room on a wagon, and (b) that room also contains a bomb. R1 seeks to extract the energy supply. In pulling out the wagon that held the energy supply, however, the bomb also hitches a ride and subsequently explodes. Even though R1 possessed the information that the bomb was in the room, and even on the wagon, just as it did in regard to the energy supply, the robot had not updated the consequences of the action of extracting the energy supply for the piece of information it had concerning the location of the bomb. This is an obvious failure to recognise the side effects of an action – i.e. to ascertain relevance. But solving this in real-time in an AI system was no easy task. At least on Dennett's telling, the heuristics that might address this problem in subsequent iterations of the robot (R2, R3, etc.) are not sufficiently "quick and dirty" to solve the problem, being stymied by computational intractability.

No doubt there are more effective bomb-locating and defusing AI systems today, with some progress concerning artificial self-other discriminations (Lanillos, 2021, 16). It does not come easily, however. By contrast it is part of the background common sense for the child, which is also enculturated with social normativity, and provides "the agent with an ability to smoothly shift from a particular context of activity into one of an open-ended number of other possible contexts of activity in a way that fits with [their] needs and interests" (Kiverstein, 2012). For Dreyfus, a key lesson is that human practical engagement with the world is not, at its most basic level, mediated by mental representations, tacit rules, or other forms of intentional content abstractable from the material context. AI systems without bodies are thus in trouble from the start, but so too are robotic bodies if designed on the GOFAI model that

[3] Birds will also "play" with the dogs chasing them, apparently deliberately remaining just out of reach.

is top-down in orientation (with CPU and rules/programs) and some other forms of AI that are bottom-up, i.e. Brooksian robotics[4]. For Dreyfus, the skilled use of a hammer is not best understood as knowing facts about how to act (cf. also Fridland, 2017) that might be stored in an AI system as a series of rules. And he thought that the relevance problem does not present in the same way for human intelligence as for GOFAI. Whilst information overload can be a problem for humans too (i.e. causing anxiety), various empirical studies suggest we can recall better when in a bodily position comparable to a situation we are trying to recall, what some researchers call the Proust effect (Morris, 2010, 238–9; Kiverstein, 2012). AI systems, by contrast, are not (yet) triggered or cued in this way, by what Dreyfus called solicitations to act and environmental affordances, which are contextually embedded. Perhaps AI systems *could* yet achieve this (Degenaar & O'Regan, 2017), perhaps even with epigenetic-like mechanisms that embed the consequences of past experiences in the 'cells' of a system, but that remains a speculative prospect (cf. Froese & Ziemke, 2009).

GOFAI researchers *did* seek to capture common sense and embodied knowledge through complex formal representations and rules and representations, using relevancy heuristics to try to avoid computational explosion. However, Dreyfus argued that these efforts won't overcome difficulties with the frame problem (Dreyfus & Dreyfus, 2000, 82). Borrowing from Ryle, he suggests that a regress of such rules of thumb looms: "if each context can be recognized only in terms of features selected as relevant and interpreted in a broader context, the AI worker is faced with a regress of contexts" (Dreyfus, 1992, 289). On his view, humans have a background know-how that derives from having bodies, interacting skilfully with material world and being trained into a culture or "form of life" of which we are often reflectively unaware (Dreyfus, 1992, xxiii). Many forms of skillful coping and engagement depend on this know-how rather than knowledge-that (facts, rules), and if we aim to embed knowledge-that into an AI system there would need to be rules about how to choose the next frame, if and when the AI system transitions between environments, or from one problem to another, and as different cognitive problems intersect in complex ways (Dreyfus, 1992, xi; Dreyfus & Dreyfus, 2000, 82, 85).

This know-how is deeply embodied across the whole organism, such that it warrants even being described as "knowledge in the hands" as Merleau-Ponty put it when talking about the skills of a pianist in his *Phenomenology of Perception*. In some domains this move seems more natural than in others, of course. But, according to Dreyfus, it is not just practical skills that it matters for, like perceiving gestalt configurations on a chess board due to training, or mastering the piano or football, say. Dreyfus extends his analyses to other kinds of reasoning, including some of those that might appear to be more "representation hungry" or "higher order" forms of cognition. While there are debates about just how far this view might extend (i.e.

[4] Dreyfus thought that Rodney Brooks' situated robotics, where the world is the best model, sidestepped the frame problem more than solved it: "Brooks's robots respond only to fixed isolable features of the environment, not to context or changing significance" (Dreyfus, 2007, 335). According to Dreyfus, the robots do not learn.

whether it plausibly includes mathematics, complex scientific theorising, etc.), for the purposes of this paper it is sufficient to compare it to basic kinds of abductive reasoning that are common-sensical (in the way Dreyfus speaks of this) rather than any more complicated theoretical inference.

Consider the kinds of abductive inference of a detective like Sherlock Holmes in Arthur Conan Doyle's work. For Holmes, in "The Adventure of the Speckled Band" say, some hypotheses are immediately rendered irrelevant or highly unlikely on physical and embodied grounds; but a hypothesis about a snake, by contrast, is apprehended as a potential explanatory hypothesis, given known facts of the case and what snakes are physically capable of (Relihan, 2009, 318). Holmes just "sees" this from perceiving the ventilation shaft. We don't all have the talent of Holmes, of course, but we know what our own and other bodies are capable of, to at least some minimal extent (based in our own habits and skills, and what we have perceived in others). This kind of abductive generation of relevant explanatory (causal) hypotheses is not so readily given to AGI systems, certainly with GOFAI assumptions and modes of operation, and arguably also in AI today.

Additionally, in empirical studies on expertise and skill-acquisition Dreyfus and Dreyfus (2000) illustrated that it is embodied perceptions of relevance, and solicitations to respond in particular ways, that are important as humans transition from being a beginner who often does learn via formal rules, through the stage of competence, and ideally eventually to expertise[5]. Their studies suggest that master chess players are scaffolded to expertise through bodily habits and Gestalt perceptions of chess figurations through embodied training, and they retain this capacity even if counting or doing other activities that block explicit reflection on chess positionality, likely moves of an opponent, etc. The general Dreyfusian claim is that it is *embodied know-how* that 'gears an organism into the world', enabling quick perceptions of relevance, habits, and basic normativity, which can be scaffolded up to something like "common sense" and prevents the frame problem's infinite regress. Again, there is some flexibility in these habits and skills, too – they are not brittle, as any mechanistic association of stimulus and response would seem to be.

How might AI systems acquire such bodies, if not through GOFAI? How is this 'relevance', this 'situated meaning', bio-physically enabled? Dreyfus doesn't give a detailed causal story about that, apart from some brief discussion of Walter Freeman's dynamic connectionism, which Dreyfus tentatively endorses due to its proposed "repertoire of attractors" not being dependent on representations (2002, 2007; see also Bruineberg & Rietveld, 2019 for useful discussion). Dreyfus was also open to reinforcement learning that is not rule-focused, and there are some interesting debates between Dreyfus and Wheeler here, which revolve around how to make Dreyfus' negative critique more of a positive agenda for AI (i.e. Wheeler, 2005). Nonetheless, Dreyfus summarises the challenge for information-theoretic computationalism in terms that remain relevant today: "unless AI scientists can produce programs in which representations of past experiences encoded in terms of salience can

[5] Note that, for Dreyfus, even the total novice presupposes a more general embodied coping as background. This is the background presupposed by the five year old, too.

directly affect the way current situations are organized, they will be stuck with some version of the frame problem" (Dreyfus & Dreyfus, 2000, 89). For him, the idea of embodied habits, situated within an "intentional arc" or a "field of affordances", is key here. Those habits extend us into the world too, and possess a historicity, in both evolutionary terms but also onto-genetically for any given individual. Can computational models (like PP) give an account of habits in this positive and intrinsically world-involving sense, given that they depend on, to at least some extent, a decoupling of representations from the world? (see, e.g. Chirimuuta, 2022). It is to that question that I now turn.

2 The frame problem and PP

In this section I will introduce some of the main ideas of PP. I will also argue that Dreyfus's discussions of the frame problem, and his contrasting account of human cognition and learning, present some ongoing difficulties for PP, despite PP depending on probabilistic computationalism rather than a deductive GOFAI model, and also not being symbol-focused with regard to its representations.

This argument is likely to appear somewhat counter-intuitive for at least two reasons. First, not many researchers talk about the frame problem today at all, whether in the quite vast PP literature, or in cognitive science, robotics, and beyond. It sometimes seems to be assumed that it is a dead problem for contemporary computationalism, including Bayesian-based PP, although not everyone has agreed (see fn 1 above). Second, we have seen some major AI achievements since Dreyfus' time. While Dreyfus confronted Deep Blue's chess success, albeit in a closed rather than open environment, we have encountered generative AI in Large Language Models (like ChatGPT) and self-driving cars (with some ongoing problems in dynamic environments), to mention but some significant recent developments. It is arguable that LLMs do now pass the Turing test, for example, depending on how much knowledge the interviewer has of LLMs. Despite this progress, mastery and expertise of certain sorts continues to elide AGI systems (Cantwell Smith, 2019), including PP based artificial cognitive systems. Are they yet capable of non-supervised intelligent and flexible behavior/cognition, with context-sensitivity? This might still be questioned, and Dreyfus' work on habits, skills, and basic abductive inference provides a potential rationale for this.

First, though, what is PP? For anybody not yet familiar, it is perhaps *the* major computational account of mind and cognition currently on offer, seeking to unify our best formal resources for understanding cognition, including the mathematical and statistical. In particular, it draws on Bayesianism inference and empirical work in predictive coding in vision and auditory science from the 1990s (i.e. Rao & Ballard, 1999; cf. also Hohwy, 2013, ch 1). Predictive coding work was primarily based on the poverty of the visual stimulus, wherein retinal imprint is compatible with different external causes, as well as with the possibility that the same external cause might give rise to different effects on sense-organs. While early GOFAI had lots of empirical problems with visual recognition as Dreyfus had emphasised (1992, 120-9), predictive coding made significant contributions to overcoming such problems,

albeit without (at this point) claiming to constitute AGIs with flexibility and context-sensitivity, or to have solved the frame problem. Rather than seek to directly represent inputs (i.e. bottom-up feature detection), predictive coding systems focus on representing prediction errors (Drayson, 2018, 3148), putting more resources into occasions where a prediction diverges from expectations, in order to avoid issues of combinatorial explosion. In this respect, however, most of these systems still depend on "seed" or "training" data, as well as human supervision and intervention, and they are not yet commonly deployed in complex real-world scenarios without supervision. If these developments are to legitimate the basic computationalist position on which the more ambitious PP program is premised, such systems should ultimately enjoy self-supervision, and exhibit greater success with the types of general problem-solving skills and flexibility that have been problematic since GOFAI. Hohwy appears to concede this (2020a, fn 14), although he and Friston and Clark seem to take PP-based AI to have been sufficiently successful to justify optimism about PP (Hohwy, 2020b, 7; cf. Friston, 2019; Clark, 2019). I think this is still an open question. This is not to say that PP-based AI has not been successful, but that in regard to AGI, which I will gloss here as general problem-solving with flexibility and context-sensitivity, the jury is out. Lanillos et al. (2021, 17) end their survey of appplications in AI and robotics with the "expectation", but it might be more accurate to say the "hope", that "the variational Bayesian inference approach will help alleviate the combinatorial explosion associated with making longer-term plans, and the accompanying deterioration in accuracy of predictions with the number of planning steps". But some researchers have noted that the frame problem has not been solved so much as by-passed (Froese & Ziemke, 2009), and it remains very difficult to engineer for context sensitivity in PP-based systems, or other versions of AGI (Kiverstein et al., 2022, 1; cf. also Millidge, 2018).

Like predictive coding, all versions of PP[6] understand the brain as fundamentally an *inference engine*, and as a *prediction machine* that uses stored knowledge and approximate Bayesian inference to predict incoming signals from the world, including the location and states of its own body. While some versions of PP are not as internalist or cognitivist as others, belief (Bayesian) and probabilistic inference are crucial to all versions of the view, with relevant algorithmic implementations and gestures towards cortical function[7]. On this view, the brain does not *passively* record inputs about the external world, then compute action plans, then seek to respond fluidly and flexibly in a dynamic and changing environment with inevitable "noise". Rather, for PP the brain is envisaged as always *actively* developing and testing hypotheses about the world. As such, PP does not have a passive conception of cognition, waiting for perceptual inputs, like early versions of computationalism and representationalism (and GOFAI). Those views were criticised by Susan Hurley

[6] In this section, I will be drawing on and discussing some of Hohwy, Clark and Friston's well-known works. I consider some other versions of PP in Section 3.

[7] This is not *necessarily* to say that physical mechanisms can be found actually performing Bayesian calculations in the brain, although it is deemed important that neurological architecture *might* physically instantiate such a system (Clark, 2013, 191; cf. Hohwy, 2020a, 17).

for their "sandwich model" of the mind (Hurley, 2001), with cognition conceived of as the intermediary between perceptual input and action output, typically with "representations" (or with 'frames', 'scripts', and 'heuristics' for GOFAI). The key PP claim is that the brain constantly anticipates upcoming sensory inputs, and creates models of its environment (representations) with the sole aim of reducing *prediction error* – that is, minimizing 'mismatch' or 'surprisal'.[8] An error is registered when the senses deliver something the brain did not expect. At that point, resources are deployed to identify the source of the mismatch in the brain's models (encoded as probability density functions) and to amend those models so that noise and error signals are minimised. These predictions are not primarily aimed at building an objective model of the world[9]. Rather, they are pragmatically oriented around controlling action and behavior that will help keep the organism viable, with a reasonably stable grip on the environment and within expected homeostatic bounds (cf. Burr, 2017; Bruineberg & Rietveld, 2019).

For PP, two fundamental principles explain both perception and action (and ultimately all cognition), with a feedback process obtaining between the two. The system *either* adjusts its model of the world to fit the inputs concerning perceptual error/mismatch, *or* it adjusts the inputs (via action) to fit the model or hypothesis that is being tested, which is to change the model via action in accordance with an agent/organism's phenotype (Clark, 2019, 3). While this might look like the "sandwich model" that Hurley criticised, the notion of the "co-evolution" of perception and action upon each other complicates this, since it makes the cognitive system less linear/serial and more dynamic, holistic, and probabilistic (which is also a computational challenge as Jerry Fodor noted long ago, a point we will return to). The precise manner in which perceptual input enables a generative "internal" model of the external world is variously construed across Hohwy, Clark, and Friston's rich works, with some versions of PP being more internalist or brain-bound than others, and with debates around the role of so-called "Markov blankets". While Hohwy seems to have a narrower conception of the relevant computations and predictions, and a more classically internal representational picture than Clark and Friston, overall I think Williams is right to emphasise some detachabilty of the model from the environment for all PP. According to him: "it is the generative model *itself* that functions as the locus of behavioural control—of the organism's active-inference induced environmental interventions—and *not* some direct coupling with the environment" (Williams, 2018, 160).

PP's focus on prediction under uncertainty as *the* fundamental mental operation enables the formal modelling of cognition as probabilistic Bayesian inference. As Hohwy puts it, Bayes' Theorem provides "a concrete sense of 'inference'", which is used to update internal models of the causes of perceptual input in the light of new

[8] This refers to that which is 'neurologically'/subpersonally surprising, rather than 'agentially' (Shannon, 1948).

[9] According to Williams, PP "posits a resemblance-based representational architecture with organism-relative contents that functions in the service of pragmatic success, not veridical representation… not linguistic or symbolic" (Williams, 2018).

evidence, in order to "arrive at new probabilistically optimal 'conclusions' about… hidden causes by weighting its prior expectations about the causes against the likelihood that the current evidence was caused by those causes" (Hohwy, 2018, 131). Hohwy's example is inferring the direction of a sound. Imagine that the system first guesses due North, receives conflicting input, adjusts, and gradually corrects to North-East. Over time, a reasoning system can settle on a stable expectation that keeps prediction error low and approximates a Bayesian inferential procedure, and without necessarily engaging in deliberate conscious inference. For PP, all organisms either act to attain some goal instrumentally (I predict I will have a coffee, and that prediction 'enslaves' the body in action until the predicted state obtains[10]), or act for epistemic reasons to increase knowledge of the environment (Hohwy, 2020b). Predictions that misfire due to ambiguous stimuli can also be 'amended' – not merely by updating the brain's model of reality, but equally by testing a hypothesis. For instance we can move towards the sound to hear it better, and the prediction error signals will get more or less ambiguous, more or less congruent with our predictions. In this respect we might think of the children's game of "Marco Polo", played with eye's shut or blindfolded. Because bodily adjustments, and epistemic actions of this nature, meet the desiderata of prediction error minimisation, PP theorists label them equally a form of inference: *active inference*. This basic picture is meant to accommodate empirical results from a range of other preceding theories[11], and to unify them via formal mathematical construals, sometimes conjoined with the idea of minimising variational free-energy, deriving from Karl Friston's work and statistical physics. The idea of minimizing surprisal (or variational free energy) might appear to require a PP system to consult potentially exhaustive knowledge of a dynamic and changing environment. But, given this is not generally possible in real-time scenarios, how exactly does the system decide and update if key probabilities are unknown to any decision-making agent, in dynamic and changing environments? What is likely, or unlikely? What set of options are even available, from which the brain might then predict likelihood and have precision expectations in regard to, thus facilitating a decision? While there is an issue of under-specification of action sequences here (cf. Burr, 2017), there are ways PP renders this idea more tractable.

2.1 Hyper-priors and the free-energy principle

First, PP theorists will generally assign a preference for organismically defined norms: relatively stable states for this or that phenotype. Evolutionary selection pressures incline an organism to congruence with its environmental niche (see Bruineberg & Rietveld, 2019), sometimes termed "hyper-priors" for a given phenotype.

[10] PP maintains that action occurs when the system prioritises a hypothesis that involves a "'systematic misrepresentation' of how our body is currently arrayed in space!" (Clark, 2019, 4; cf. Hohwy, 2016, 276). Action is thus akin to a self-fulfilling prophecy (Friston, 2009, 295).

[11] For example, it is not the "pleasure principle" (i.e. reward) that is envisaged as directly causing behavior, but the minimizing of prediction error and (expected) rates of change of prediction error, which establishes valence (Miller et al., 2020; Clark, 2019), with dopamine systems a key mechanism.

As such, surprisal is both phenotypically relative, as well as onto-genetically relative via 'priors'. Nonetheless, the PP view is that it stills warrant an inferential and predictive construal, because the species as a whole can control this over longer evolutionary time frames, due to new habits and habitats (with niche construction) and selection pressures that are responded to. Evolution is the minimising of surprisal at phylogenetic timescales (Sims, 2017, 7).[12] According to Friston's concept of the Free Energy Principle (FEP), prediction 'errors' are interpreted as departures from an organism's homeostasis, which trigger attempts to return to preferred set points (Friston, 2013, 2019). Hohwy's rendering of the FEP requires that the organism posit a principled boundary between itself and the rest of the world (Hohwy, 2020a). This is computationally modelled via a Markov blanket, which defines the boundaries of a system in a statistical sense, separating the states that make that thing the particular kind of thing that it is, from the "external" states that it is not (Parr & Friston, 2019). It also relies on the assumption of ergodic density, an invariant probability measure pertaining to likely states of the organism. The FEP hence serves as a mathematical and statistical framework that is applied to all organisms that must solve the problem of continuing to exist: i.e. resisting entropy and the second law of thermodynamics for a period, maintaining structural integrity rather than dissipating, and thus attaining a non-equilibrium steady-state (cf. Parr & Friston, 2019). Again, this enables approximate Bayesian inference, since variables can be specified in relation to a given phenotype and its meta-stable states via ergodicity. It remains, however, an information-theoretic way of modelling uncertainty reduction, applying to all biological systems but based on computational functions and construing value and fitness in terms of belief (uncertainty reduction).

2.2 Precision-weighting

Another important part of the PP account of cognition is the idea that the system can assign different *precision weightings* to incoming evidence, given what it already knows (its 'priors'). To refer back to the earlier example, a sound-identifying system might initially guess due North because that is the usual direction sounds come from in their environmental niche, whether based on direct experiences, or longer-term evolutionary imperatives that are phenotypically embodied. PP systems hence model their own *precision* in modelling the world, which enables strongly-held or high-precision beliefs and expectations to be preferentially defended. The PP account of modelling precision hence involves a hierarchical generative model of the brain's attempts to minimize prediction error on a number of levels simultaneously, which Clark describes as a "cascade of cortical processing events in which higher-level systems attempt to predict the inputs to lower-level ones on the basis of their own emerging models of the causal structure of the world (i.e., the signal source)" (Clark, 2013, 181-2). Clark says this allows for an "astonishingly fluid and

[12] Of course, we might wonder just how chance mutation, ostensibly a force in evolution even on gradualist views, is adequately understood as a prediction.

context-responsive" system (Clark, 2018, 523), perhaps even a cognitive system for which the frame problem might not present, if artificially simulated.

But we need to consider what is doing the work here – the formal and probabilistic architecture of the PP tool-box (i.e. the computational account), or other features that are smuggled in by starting with biological organisms and their already acknowledged context-sensitivity. PP takes what we already recognise as adaptive and living organisms and asserts a mathematical/functional process (which is substrate neutral) that is necessary for such survival, but there are no falsifiable examples and little explanatory power in terms of explaining any particular forms of biological cognition. It has been argued that this is because PP trades off biological plausibility and specificity for general applicability via probability and statistical physics, and some have argued that it elides the differences between organismic robustness and homeostasis (Colombo & Palacios, 2021; cf. also Litwin & Minkowski 2020/22). Does PP's reduction of uncertainty (surprisal) solve the problem of when and where to exploit/explore when foraging in the 'wild' say, and a choice between pragmatic and epistemic action? Sophisticated biological organisms also need to be able to switch between such actions, and decide between at least the basic evolutionary imperatives sometimes called the "4 fs": fighting, fleeing, feeding, fornicating (and they also sleep, idle, play, etc.). It remains in question precisely how the high-level imperative to minimise surprisal or "free energy" enables an array of available choices and action-possibilities that will be probabilistically sorted in terms of prediction error minimisation, and precision-weighting.

And to return to the basic dilemma for computationalism, which PP remains a form of, if the *raison d'etre* of cognitive systems is to minimise prediction error and to generate complex hierarchical (or dynamic) models regarding prediction errors that approximate to Bayesian rules, then it seems that PP based AI not only could do that very effectively but would readily surpass humans at it, i.e. at computations over probability distributions. But it is not clear that PP or "active inference" robotics have (yet) emulated or surpassed the flexibility of human problem-solving in dynamic contexts (for surveys, see Millidge, 2018 and Lanillos, 2021), nor the kinds of embodied habits and inferences we sketched in the above account of Dreyfus. We seem entitled to wonder why not. Given the unifying and imperialistic claims of most versions of PP – it is often said to be the sole or fundamental cognitive principle (cf. Clark, 2019, 1, Howhy, 2015) – a full response to the frame problem that addresses the apparent differences between artificial and biological cognitive systems seems needed. While this has not yet been convincingly provided by PP, to my eyes, there are indications that some think that it has been provided already, either explicitly or implicitly.

2.3 PP on the frame problem: explicit and implicit

Of those who appear to think the frame problem has already been answered, Linson et al. (2018, 15) claim that it is just a problem for the "in-output" model of cognition, and that the Active Inference framework "dissolves" the frame problem. In particular, Linson et al hold the problem of relevance is addressed via nested priors,

and that the more technical *logical* frame problem (see fn 2) is obviated by the probability distributions of the generative model. While this move indicates the contours of a possible solution, fully mapping any organism's priors (everything? If not, which priors?) and hyper-priors (via a statistical average? But species do change their key habits and sometimes quickly) seems an enormous task. In addition, questions remain regarding whether it is not essentially a redescription, and the extent to which the modelling fully explains the target (flexible, context-sensitive beahvior, grounded in habits and skills). Intelligent biological behavior is posited as involving nested priors, which seems plausible, but formally modelling this in dynamic and real-world scenarios, or building AGI systems on this basis, remains very much work in progress, despite all of the available resources for probabilistically minimising prediction-error. Without further progress, we are within our rights to be agnostic about claims the problem has been dissolved, and that it has dispensed with Dreyfusian style worries.

Furthermore, the claim that the logical frame problem has been obviated hints at a broader point that needs to be addressed by the Bayesian-based PP and active inference literature. Linson et al.'s (2018) purported solution is Bayesian (as is Hohwy's), but Bayesianism is generally not seen as a solution to the frame problem, but rather a procedure for decision in conditions of uncertainty. Once you have certain information – priors, or likelihoods – you can reason effectively, despite uncertainty. But you still need some input, so the generation of these needs to be accounted for, as well as why (and 'how') this or that set of possible hypotheses is generated, rather than indefinitely many others. We have seen these are imported into PP via the assumption of invariant ergodic states (hyper-priors), and 'priors' pertaining to the life-history of this or that organism. The merits of the former move are debated in philosophy of biology, as we briefly saw (cf. Litwin & Milkowski, 2020; Palacios & Colombo, 2021).

But there is a deeper problem. Even if we admit such a move, idealising as it is, basic abductive reasoning of the kind I have associated with Dreyfus above, including in the Sherlock Holmes example, is not Bayesianism simpliciter, or at least it is not obviously so. Creative abduction involves the generation of hypotheses, but Bayesian decision theory works with the likelihood of an existing range of hypotheses. No doubt the claim is intended to be that PP combines both: active inference as a kind of creative abduction in C.S. Peirce's terms, and then Bayesian decision procedures to resolve uncertainty[13]. There is something to this idea, as applied to embodied and biological organisms, but it is not clear how to mathematically formalise both procedures, nor precisely how they might interact, given an organism holistically interacting with its environment, with a set of habits and skills. Even if we restrict ourselves to the context of selection from a range of hypotheses, what is

[13] Linson, Schulkin, Clark have a 2022 paper where they cite Friston (2018) and treat Peircian abduction as a "form of Bayesian model selection". Later in the same paper they treat Bayesian or "abductive optimality" as equivalent. But abduction is not (at least for Peirce) about optimality at all. For more on this, see Legg, 2001.

referred to in the literature as "selective abduction", the results of such abductive reasoning still may not precisely align with Bayesian techniques (see Douven, 2022).

In general, abduction and Bayesianism have a complex inter-relation that is rarely acknowledged in the PP literature. Some epistemologists and philosophers of science think that they are compatible, but many others maintain that they are opposing views. In either case, they are not the same thing, even for Lipton (2003) and those who attempt to conjoin them. *Prima facie*, there is a difference between degrees of belief versus categorial belief, and computationally modelling abduction (unlike Bayes' rule) is argued by some to be inordinately difficult, perhaps intractable (Douven, 2022, 15; Kwisthout et al., 2011). Yet Hohwy[14], Friston and others seem to assume that abductive and Bayesian inference are identical, but Van Fraassen and many others have thought that this is not so (Douven, 2022). Indeed, Van Fraassen dismisses abduction for that reason: for him (and Bayesians more generally) any systematic way for changing belief that is not Bayesian is not rational. But abduction, for Douven, involves "a bit of art, of imagination, and of creativity" (Douven, 2022, 25). Moreover, abduction is context-sensitive and relies on background knowledge (akin to what Dreyfus calls common sense). Douven notes that abduction fits better with an ecological conception of rationality than Bayesianism, which on his portrayal is more one size fits all. Abduction is specific to an organism or reasoning agent and their particular circumstances, admittedly in the way that at least some proponents of PP also want to embrace (see Section 3 below). While I cannot definitively settle the question of the relation between abductive and Bayesian inference (and IBE) here, I have given reason to question any simple equivalence, and this matters for my ensuing argument, since embodied inferences and habits of many biological organisms may also be better understood abductively rather than through Bayes' rule, or as involving both, but in a way that may be computationally intractable in dynamic real-world scenarios.

Although Hohwy does not say this explicitly, there are some indications that he thinks PP has adequately addressed the problems that are often characterised in terms of the frame problem (i.e. relevance, meaning), through *precision-weighting*. For example, he says:

> …hierarchical precision-weighted predictive coding is critical because predictive coding then becomes context-sensitive, enabling it to deal with irreducible noise in the input, ambiguity in the internal model's mapping of hidden causes to expected sensory input, and volatility in the sensory input due to interactions between hidden causes (Hohwy, 2020a, 3).

Hohwy also appears to think that such precision-weighting enables genuinely *self-supervised learning*, as "Systems that can minimise error *only* need to

[14] While Hohwy (2013) does not refer specifically to abduction, he does discuss something that is sometimes thought to be abduction, that being inference to the best explanation (IBE). He also claims the inference type, IBE, is essentially Bayesian (2013, 25). Few people hold this. Even Lipton (who Hohwy cites) does not, although he insists they are compatible and mutually complementary. Douven argues against this, and also notes that Hohwy does not present arguments for the connection (2022, 5).

access their model and the sensory input." (Hohwy, 2020a, 3, my italics; cf. also Hohwy, 2020b). There are two points that warrant attention here, however. The first concerns whether Hohwy's fairly minimal conditions for prediction-error mechanisms are sufficient for self-supervision. The second and more crucial point concerns what else might be required to account for the fully flexible cognition that Dreyfus argued was characteristic of human skills and expertise, and grounded in the body, which he maintained constrained the space of possible generalisations, enabling perceptions of relevance, trial and error learning, and basic abductive inference (which is not reducible to Bayesianism, or at least not yet shown to be so).

On the question of self-supervision, recall that Dreyfus and Dennett highlighted the problems associated with embedding deixis in GOFAI systems. In Dennett's dramatic tale, the bomb ended up exploding, yet it seems that R1 (and R2, R3, etc.,) possessed Hohwy's prerequisites. That is, R1 could access its model and its sensory input, but it did not end well. Of course, R1 (or R2, R3, etc.) was not a PP based model of AGI. Lanillos et al. (2021) suggest some improvement here, but proprioception and context-sensitivity in regard to self and other induced movements in contemporary AI remains distant from that exhibited by many biological organisms. While Hohwy argues against the importance of body-ownership for cognition (2016, 115, cf. also Hohwy & Michael, 2016), and also claims that "organisms do not need to have the capacity to distinguish between self and other to function appropriately" (2016, 116), both of these claims are contentious and it is not entirely clear on the nature of their empirical support. In particular, some form of self-other distinction seems necessary for organisms to attain a certain level of cognitive sophistication (cf. Godfrey-Smith, 2020), including flexible intelligence. Without it, a cognitive system's own motions will confound its efforts to understand what is going on in the world (e.g. the crab might grab its own claws and treat them as a danger). Reinforcement learning that is context-sensitive and flexible appears to be based on indexical self-hood that differentiates self-produced from externally caused action, and helps to determine relevance. This returns us to our claims regarding the important role of embodiment in assisting with basic abductive inferences regarding relevance, and ultimately intelligent behaviour and cognition.

While Clark and others have sometimes resisted Hohwy's more disembodied view of PP (see Section 3 below), my basic claim here is that if Clark and company are invested in PP's self-sufficiency (its cognitive imperialism), as they also sometimes seem to be, then a response is required to what Fodor called the "riddle of abduction" for computationalism: that is, whether the computational cognitive system might simultaneously be physical, reliable and feasible (Relihan, 2009). One way of solving the problem, of course, is to treat abduction as identical with Bayesianism, which enables precise formal algorithms (which AI would be better placed to deliver on than humans). This is the route that PP theorists have taken, but without showing this identity formally (Douven, 2022) and without (yet) delivering flexible and context-sensitive AGI. So, there is work to be done by them in this regard. And to come to Fodor's trilemma, so far the jury is out on feasibility and physical implementation in PP-based AGI systems in the real world (computational modelling may be another matter). It is also not easy to see how PP and active inference views can have recourse to one of the key features of that computational triad discussed by

Fodor: information encapsulation, without which the system may never stop thinking, thus confronting Hamlet's problem. PP's generative models and mechanisms are not isolated or homuncular, but holistic. They may do better regarding relevance *because* they are not informationally encapsulated, but therein also lies the computational challenge in terms of feasibility.

2.4 Being, doing, and habits

With regard to possible PP/active inference implementation in AGI systems, one difficulty is the absence of a strong or "intrinsic" relationship between "being" and "doing" for such systems, to invoke some points made by Froese and Taguchi (2019). By contrast, a biological organism's precarious existence (being) depends on what it does, including the flexibility or otherwise of its habits, and the consequences of particular choices are not formal but material, even epigenetically inherited in particular cells in particular bodies (Meloni & Reynolds, 2021). Would a robotic body help? Not according to Cantwell Smith at least under current second and third wave AI (2019), and not according to Dreyfus, whose work suggests that a system's access to its own generative model and sensory input is necessary but not sufficient for intelligence. It would need to be a body that encoded the past in habits, and which were also open to particular futures, cuing organisms for context and relevance, scaffolding learning, and enabling flexible problem-solving in situations of complexity. The robotic body would require indexicality and "deixis" in the sense of Dreyfus' five year old child, not just access to a "model" or "map" that might be updated in the light of new evidence, but being able to register the difference between the real world and the model that stands for/represents it, which indexicality and self-other discriminations are vital for. Cantwell Smith suggests this is an important part of the distinction between mere "reckoning" (or calculation) and intelligent judgment (2019).

There is also a feedback loop between being and doing, in which an organism's existence – metabolically, homeostatically – depends on its own doings and decisions. As Kiverstein et al. (2022) put the point: "It is this relationship between being and doing that makes for goals and concerns that are intrinsic to the organism". In regard to the AI implementation of such frameworks, Millidge notes the computational difficulties that confront an active inference framework that takes seriously the synergistic relationship between being and doing (and simultaneous model revision and action in the world). Millidge suggests that some of those systems have minimised "prediction error through a compromise of action and sensory updating", but in a way that has been "deleterious for the agent", which he has attributed to the system doing both at the same time but thereby compromising both the robot's perception and action models.

This feedback relationship between being and doing involves what Raja et al. (2021) call "relational features". They argue that these are characteristic of any system with sufficiently dynamic and fluid relationships between its parts, and between that which is putatively internal and external to the system. For them, it includes the skills, habits and "solicitations to act" that Dreyfus emphasised, along with the

"motor intentionality" that enables the relational understanding of "nearer than", "larger than", "smaller than", "closer to", and affordances, with their embeddedness in specific contexts. These intrinsically relational features pose a computational challenge. The question is whether they are adequately understood via an internalist, statistical model, which PP develops through the formalism of Markov blankets and probability distributions with precision weighting and action policies, based on identifying stable states in the system that could correspond to blanket states (Raja et al., 2021). As we saw with the postulation of ergodic densities above, this move separates the elements of the structural coupling, the organism and the environment, but in breaking habits down in this way the relationality itself appears to be attentuated. Modelling it via inner or outer states of the system may be instrumentally useful, of course, but it is not yet clear it can capture the dynamics of habits and skills without remainder, pending developments in PP-based AI and robotics.

Recall also that Dreyfus has argued that we encounter more and more refined "solicitations to act", as expertise increases. That insight is accommodated by PP, to some extent, since agents act in order to minimize surprise about their own future states. This future-orientation means that the consequences of past perceptions and actions becomes part of one's 'priors', along with inferences regarding the likelihood of change/stability. An account of habits as the selection of those actions and policies that are least likely to induce "surprisal" also follows (see Miller et al., 2020). In general, however, the portrait of habits is quite variable in PP and the Active Inference framework, perhaps reflecting uncertainty about just how such features are to be best understood. Are they intrinsically relational, as Dreyfus holds, or not? If they are intrinsically relational, with deep coupling between organism and environment, then any computational treatment of them will be, to say the least, exceedingly complicated.

Instead, some PP proponents have rather reductive takes on habits[15]. For example, Linson et al. suggest that "for AIF, habits can be regarded as context-free responses that are established by their invariance across multiple conditions. When we act out of habit, we merely go through the motions, suppressing any potential significance that might otherwise be contextually relevant" (2018, 15). This is not an uncommon claim about habits in cognitive science and AI, which are both still generally computational in outlook. However, it misses the Dreyfusian challenge about the nature of human (and animal) habits. The master or skilled expert is still acting in what he calls *l'habitude*. This is part of what allows them to establish relevance and attend with precision, but it is not brittle. And it is situated within a broader field of affordances (Bruineberg & Rietveld, 2019). In short, habit is not the opposite of innovation on Dreyfus' view, but a key condition of it. Without that worldly embeddedness, orientation, and associated network of meaning that is a practical series of

[15] Others are less reductive. Froese, Bruineberg and Kiverstein have a better view, for example, and they want to embed habits in forms of life, borrowing from Wittgenstein, to address the problem of meaning. I come back to their work in Section 3 below. I claim they might be taken to question PP's explanatory imperialism, with which I agree. As far as I can see the relationship between abductive reasoning and strict Bayesianism is not resolved: while the latter is computationally tractable the former is less so.

possibilities and potential actions and consequences (i.e. what Dreyfus and Merleau-Ponty call an "intentional arc"), the question is whether any PP-based AI systems might be said to genuinely have skills and habits in Dreyfus' sense, the grounds of flexible intelligence for him. While a more mechanistic rendering of habit is much more computationally tractable, any automaticity of habits with one-to-one stimulus-response pairings counts *against* flexibility and context-switching (i.e. general intelligence). Moreover, that is not congruent with the details of Dreyfus' empirical studies of human skill acquisition and expertise, as presented in Section 1.

Dreyfus's treatment of the habits/skills nexus also highlights the concrete contextual affordances that present as part of ordinary visual experience. As Kathryn Nave nicely puts the point, aiming to contest some aspects of Clark's understanding that experience-based selection is concerned only with generic action types:

> The kind of actions that visual experience allows us to select between are not just abstract types like 'eat' vs. 'throw', but rather the particular sort of throws that would, or would not, achieve my goal, given my current bodily position and environmental situation (Nave, 2022, 404).

While Nave remains broadly supportive of PP, these concrete action-possibilities are relative to particular organisms and their particular environmental niche, as well as relative to specific morphological constraints. The specificity of these bodies places pressure on a computationalist construal, especially any computationalism that – like PP – claims to be the sole cognitive principle. On Nave's account, they are a fundamental part of our perceptual orientation in a world, much as Dreyfus' claims about the "direct perception of significance" (2002) are central to his arguments about why master chess players are able to retain their level of skill, despite being experimentally forced to respond very quickly with each of their moves and while also doing other cognitive activities at the same time (Dreyfus & Dreyfus, 2000). Are these contextual and embodied responses adequately understood as "action policies", as PP contends? Action policies are both computational and generic, rather than specific (and directly relational) in the way that environmental affordances solicit particular actions in particular ways, and for which the more "ecological rationality" of abduction seems a better fit. Notice also that Clark's emphasis on generic action policies does not seem far away from "relevancy heuristics", thus potentially encountering some of the problems concerning infinite regress that Dreyfus associated with the frame problem. Which action policy? How and when might these be switched in novel circumstances? We do not yet seem to have overcome the problem of under-specification of actions, or established in detail the generation of the range of options upon which PP's Bayesian probabilistic inferences to minimise prediction error might work. The generation of available hypotheses is not itself explained, other than via hyper-priors and homeostatic norms in intelligent biological organisms, and there remains a lot of different ways to stay within homeostatic norms and to avoid dissipation/death.

In sum, PP has not yet adequately addressed the frame problem, nor fully grappled with the apparent differences between biological and artificial intelligence, many of which Dreyfus had highlighted in the 70s and subsequently. There is a plausible argument, building on that body of work, that genuinely flexible cognition

depends on deep organism-environment relationality, grounded in embodiment and the sense of deixis it provides (also see Di Paolo et al., 2022). Embodied skills and habits, in systems with intrinsic temporality and deep relationality, are amongst the key factors in biological intelligence (Piersma & Gils, 2011, 126-7; cf. also Godfrey-Smith, 2016) that exhibits the kind of relevance-based context-sensitivity and flexibility that is of interest to Dreyfus.

3 Replies to some objections: pluralism; relevance; and the body as the model

In what follows I outline some potential objections to my "framing" of the predictive mind. These concern: the *pluralism* of some versions of PP; the idea that the *embodied organism is itself the model*; and the claim that *relevance just is precision-weighting*.

I noted above that Clark sometimes seems to advocate a more pluralist version of PP, reaching out to rival proponents of embodied cognition like "enactivism" who can fill in details not provided through PP's predictive formalism. They might even act as a "perfect partner" (Clark, 2016, cf. Clark, 2013, § 5.2). While strong claims regarding the explanatory sufficiency of PP remain the orthodoxy, I must concede that I find Clark tricky to interpret on his commitments in this regard, as well as on the question of whether or not he is proposing a pure or hybrid version of PP, given his criticisms of body-chauvinism and his continued use of generic action policies (Clark 2007, Clark 2008; but. c.f Anderson, 2017; Nave, 2022). Clark sometimes admits difficulties of the kind I have drawn attention to, noting that "selecting which action to perform next, given a large body of world-knowledge, is computationally challenging. *But it is no more challenging*, using PP-resources, than it is, using more traditional ones" (Clark, 2019, 7, my italics). Clark does not say the problem of action-selection is *less* challenging using PP resources, just that it is not *more* challenging. With this weaker language there is perhaps not the same implication in some of Clark's work, as compared to Linson et al., 2018 (although Clark is a co-author), or Hohwy's work, that PP has overcome the frame problem. But any claim that PP's formal Bayesian model explains *all* mind and cognition, both biological and with extension to artificial systems, is also potentially placed in question.

Without seeking a definitive conclusion about Clark and the extent of his pluralism here, there are a range of other views of active inference within PP that also endeavour to take the holistic body more seriously, often directly indebted to Friston and the FEP. This includes Bruineberg et al. (2018), who have a more ecological-enactive version of PP, wherein active inference is understood to pertain to the whole situated organism (not merely the brain), within bounds prescribed by its phenotype. Most relevantly, however, Kiverstein et al. (2022) recently published a paper on the problem of meaning, bearing directly on the frame problem. In brief, they conclude that the FEP needs enactive cognitive science for the solution it provides in regard to meaning (2022, 3). I agree with this claim, but to me it suggests that the FEP by itself cannot account for the frame problem: it needs supplementation. They also accept that most versions of "active inference" have pre-specified the search space

(Kiverstein et al., 2022, 8), thus addressing the frame problem from outside. And they also hold that sensorimotor autonomy makes a crucial difference to problem of meaning (2022, 8), contrary to the account from Hohwy sketched above, a point with which I also concur. As such, it is arguable that they do not endorse strong computationalism – i.e. the mind/cognition is essentially (or nothing but) computations, in this case to minimise prediction error – or the idea that an information-theoretic construal is alone sufficient. If that is so, I take their work to be broadly compatible with my arguments here. I also think that they have the right views about habits, although I am less optimistic that modelling past experiences as priors, and future orientation understood in terms of precision estimation, will suffice. As noted above, any mapping of relevant priors and hyper-priors in complex organisms is already an enormous task. Although habits (and skills) are indeed forms of uncertainty reduction admittedly, the inferences that embodied organisms make might be abductive more than Bayesian, and it is not clear how to resolve formally this difference, given the dynamic and continuous reciprocal causation between organism and environment involved. It can be modelled as such, but we don't want to mistake the model for the target, and until we have PP based AGI with flexible problem solving I think we should remain agnostic about this.

Some other recent proponents of what we might label a "PP+" approach appear to treat precision-weighting as identical with relevance, notably Andersen et al. (2022). We saw that Hohwy implied such an identification above too. *If* that identity holds good, then it might solve the frame problem too, or at least make substantial headway towards that goal. Is relevance nothing but precision weighting, with a high precision weighting assigned to relevance, and a lower precision-weighting assigned to the irrelevant? But doesn't this beg the question? Does it *explain* relevance, or just insist that PP's precision weighting *is* relevance? Precision weighting might facilitate a human or artificial system to drive a car in situations of poor light, or unknown roads, compared with in good light or in known environments. But it is, of course, when the genuinely novel enters the picture that problems can occur, especially with driverless cars in cities. Consider the difficult to predict behavior of pedestrians, kangaroos, cows, people, bikes, just to hint at your average city! Saying that precision weighting is relevance does not dispense with the algorithmic complexity.

In addition, Andersen et al. (2022) say that the two frameworks in question, PP and Relevance Realisation, "mutually support". That may be so, but it suggests that PP needs the relevance realisation framework and is not sufficient by itself. If so, this appears to support my argument. Their idea of "mutual support" seems to need further specification, however, since it is neither causally explanatory of either, nor a strict claim of logical identity. If we look at the details of the Relevance Realisation framework they propose to add to PP, they discuss a series of trade offs that are required by an organism. They discuss the efficiency-resiliency tradeoff, as well as three derivative tradeoffs, which are the exploration-exploitation tradeoff, the specialization-generalization tradeoff, and the focusing-diversifying tradeoff (Andersen et al., 2022). They argue that these trade-off relationships constrain what an organism can find relevant. While I agree, these broad parameters don't seem to help alot with regard to generating particular choice options, nor deciding between particular actions that might be good exploratory options, or good exploitative options.

These are presumably many more than just two, in real-world scenarios. What cues are suggestive of focusing and what options for focussing are there, what cues are suggestive of diversifying and what are the available options in that regard? How exactly do the three trade-off situations determine an action or decision? How do the high-level constraints that these exercise on each other determine how we perceive a situation as warranting a switch in strategy, say, which is what is needed for flexible intelligence (and AGIs)? In brief, my concern is indeed that "precision weighting" becomes akin to a magic modulator.

Finally, while they rightly give an important role to affect in securing context-sensitive precising weighting, which is clearly an important factor in the situated normativity of biological intelligence (cf. Kiverstein et al., 2022), Andersen et al. (2022) also counterpose habits as repetitive behavior (akin to exploitation) rather than being a form of discovery or exploration. But, one of the key Drefyusian claims is that that habits have some flexibility built into them, enabling coping with always slightly different environments and contexts, and to be the basis from which skills can develop. As such, habits are updating and intelligent for Dreyfus and are revised over time, with embodied know-how. We can have bad habits, of course, which close us off to opportunities (perhaps forms of compulsive addiction, say), but these are one possibility of habits that do not define their nature.

Sometimes this more body-centric view is clear in other PP work too, but it sits uneasily with some of PP's basic formal and computationalist commitments. For example, in filling in the role of evolution in regard to priors, Parr, Pezzulo and Friston (2022, 4) note that: "This implies that the generative model of each particular animal is tightly constrained by the statistics of their ecological niches and the control demands of their bodies". In seemingly referring to particular bodies, not just generic action policies and phenotypes with invariant ergodic states, has a form of "body chauvinism", where particular bodies are special, returned to PP? Indeed, Parr et al. also state that all brains have "*some* predictive motifs" (my italics, 2022, 4). But some is not all, perhaps indicating they are averting here to pluralism, or hybrid versions of PP. If so, cognition does not seem to be reducible to PP's computational tool-box without remainder. While I agree with this position, proponents of PP face a risk with such a move: notably the unifying claim that prediction error minimisation is the sole cognitive principle. At the very least, such holistic views of the body/organism as model concede quite a lot to the view I have put forward here, even if they do so within the terms of an expanded and adjusted PP: PP+, or hybrid PP. In addition, and with regard to the frame problem, they have not resolved the formal differences between abduction and Bayesianism discussed in Section 2.

4 Conclusion

I have suggested that PP confronts an issue concerning whether it can both explain all cognition and intelligent behavior as minimizing prediction error with just the basic formal and computational elements of the PP toolbox, *and* adequately comprehend the prima facie differences between cognitive systems such as biological and artificial intelligences. Dreyfus's treatment of the frame problem, and the role

he gives to embodiment in resolving it, outlines the difficulty of the challenge, since for him there is a deep connection between embodied indexicality and the ability to develop habits and skills in interaction with an environment, and hence flexible and context-sensitive biological intelligence, including basic abductive inference. Recent PP modelling of these aspects does not formally consider the differences between abduction and Bayesianism, nor the full complexity of Dreyfus' more positive and dynamic construal of habits[16]. While predictive mechanisms are clearly an important part of understanding cognition, my argument has been against the self-sufficiency of PP's Bayesian and computational framework. Justifying PP's more ambitious claims requires an explanation about why context-sensitivity and flexible general intelligence remains difficult for formal models of cognition utilizing PP, and why it remains more characteristic of biological than artificial intelligence. PP theorists either need to provide a formal and computational answer to the frame problem that makes a difference to contemporary robotics and AI (and addresses the relationship between abduction and Bayesianism), or else they need to moderate some of their stronger rhetorical claims regarding being the sole cognitive kind.

Acknowledgements I am indebted to Cathy Legg for our early conversations around AI and predictive processing, as part of what was initially envisaged as a co-authored paper. More recently, the essay has benefitted from invaluable feedback from Julian Kiverstein, as well as a number of peer reviewers for *PCS* that have all helped me improve the manuscript. Finally, Ross Pain and James Williams also offered useful feedback a year or two ago, for which I am grateful.

Author contributions NA.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability NA.

Declarations

Ethical approval NA.

Informed consent NA.

Statement regarding research involving human participants and/or animals NA.

Competing interests NA.

[16] Except Kiverstein et al. (2022), cf. fn 15.

References

Andersen, B. P., Miller, M., & Vervaeke, J. (2022). Predictive processing and relevance realization: Exploring convergent solutions to the frame problem. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-022-09850

Anderson, M. L. (2017). Of bayes and bullets: An embodied, situated, targeting-based account of predictive processing. In T. Metzinger & W. Wiese (Eds.), philosophy and predictive processing (p. 4). Frankfurt am Main: MIND Group. https://doi.org/10.15502/978395857305

Bruineberg, J., & Rietveld, E. (2014). Self-organization, free energy minimization, and optimal grip on a field of affordances. *Frontiers in Human Neuroscience, 8*, 599. https://doi.org/10.3389/fnhum.2014.00599

Bruineberg, J., & Rietveld, E. (2019). What's inside your head once you've figured out what your head's inside of. *Ecological Psychology, 31*(3), 198–217. https://doi.org/10.1080/10407413.2019.1615204

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: The free energy principle from an ecological-enactive perspective. *Synthese, 195*, 2417–2444. https://doi.org/10.1007/s11229-016-1239-1

Burr, C. (2017). Embodied decisions and the predictive brain. In T. Metzinger, & W. Wiese (Eds.), *Philosophy and Predictive Processing: 7*. MIND Group. https://doi.org/10.15502/9783958573086

Cantwell Smith, B. (2019). *The promise of artificial intelligence: Reckoning and judgement*. MIT Press.

Chirimuuta, M. (2022). Disjunctivism and Cartesian idealisation. *Proceedings of the Aristotelean Society, 122*(3), 218–238.

Clark, A. (2007). Reinventing ourselves: The plasticity of embodiment, sensing, and mind. *Journal of Philosophy and Medicine, 32*(3), 263–282.

Clark, A. (2008). Pressing the flesh: Exploring a tension in the study of the embodied, embedded mind. *Philosophy and Phenomenological Research, 76*(1), 37–59.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioural and Brain Sciences, 36*, 181–253.

Clark, A. (2016). *Surfing uncertainty*. Oxford University Press.

Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomena Cognitive Science, 17*, 521–534. https://doi.org/10.1007/s11097-017-9525-z

Clark, A. (2019). Beyond desire. Agency, choice and the predictive mind. *Australasian Journal of Philosophy*. https://doi.org/10.1080/00048402.2019.1602661

Colombo, M., & Palacios, P. (2021). Non-equilibrium thermodynamics and the free energy principle in biology. *Biology and Philosophy, 36*, 41. https://doi.org/10.1007/s10539-021-09818-x

Dennett, D. C. (2006). The Frame Problem of AI. *Philosophy of Psychology: Contemporary Readings BOOK*, *433*, 67–83.

Degenaar, J. & O'Regan, K. (2017). Sensorimotor Theory and Enactivism. *Topoi, 36*(3), 393–407. https://ejap.louisiana.edu/ejap/1996.spring/dreyfus.1996.spring.html

Di Paolo, E. A., Thompson, E., & Beer, R. D. (2022). Laying down a forking path: Tensions between enaction and the free energy principle. *Philosophy and the Mind Sciences, 3*. https://doi.org/10.33735/phimisci.2022.9187

Douven, I. (2022). *The art of abduction*. MIT Press.

Drayson, Z. (2018). Direct perception and the predictive mind. *Philosophical Studies, 175*, 3145–3164.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT Press.

Dreyfus, H. L. (2002). Intelligence without representation: Merleau-Ponty's critique of mental representation. *Phenomenology and the Cognitive Sciences, 1*(4), 413–425.

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

Dreyfus, H., & Dreyfus, S. E. (2000). *Mind over machine*. Simon and Schuster.

Dreyfus, H. (1998). The Current Relevance of Merleau-Ponty's Phenomenology of Embodiment. *Electronic Journal of Analytic Philosophy*. https://ejap.louisiana.edu/ejap/1996.spring/dreyfus.1996.spring.html

Fridland, E. (2017). Skill and motor control: Intelligence all the way down. *Philosophical. Studies, 174*(6), 1539–1560.
Friston, K. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13/7*, 293–301.
Friston, K. (2013). Life as we know it. *Journal of the Royal Society Interface / the Royal Society, 10*(86). https://doi.org/10.1098/rsif.2013.0475
Friston, K. (2018). Does predictive coding have a future? *Nature Neuroscience, 21*, 1019–1021.
Friston, K. (2019). A Free Energy Principle for a Particular Physics. arXiv [q-bio.NC]. arXiv. http://arxiv.org/abs/1906.10184. Accessed 1 January 2023
Froese, T., & Taguchi, S. (2019). The problem of meaning in AI and robotics: Still with us after all these years. *Philosophies, 4*(2), 1–14.
Froese, T., & Ziemke, T. (2009). Enactive artificial intelligence. *Artificial Intelligence, 173*, 466–500.
Godfrey-Smith, P. (2020). *Metazoa*. William Collins.
Hohwy, J. (2013). *The predictive mind*. Oxford University Press.
Hohwy, J. (2015). Prediction, agency and body ownership. In A. Engel, K. J. Friston, & D. Kragic (Eds.), *The pragmatic turn: Toward action-oriented views in cognitive science* (pp. 109–120). MIT Press.
Hohwy, J. (2016). The self-evidencing brain. *Noûs*. https://doi.org/10.1111/nous.12062
Hohwy, J. (2018). The predictive processing hypothesis. *The Oxford handbook of 4E cognition* (pp. 129–146). Oxford University Press.
Hohwy, J. (2020a). New directions in predictive processing. *Mind & Language, 35*(2), 209–223. https://doi.org/10.1111/mila.12281
Hohwy, J. (2020b). Self-supervision, normativity and the free energy principle. *Synthese, 199*, 29–53. https://doi.org/10.1007/s11229-020-02622-2
Hohwy, J., & Michael, J. (2016). Why should any body have a self? In F. Vignemont, & A. Alsmith (Eds.), *The subject's matter: Self-consciousness and the body*. MIT Press.
Hohwy, J. (2016). Prediction, agency, and body ownership. In K. Andreas, K. Engel, J. Friston, & D. Kragic (Eds.), *The pragmatic turn: Toward action-oriented views in cognitive science* online edn, MIT Press Scholarship. https://doi.org/10.7551/mitpress/9780262034326.003.0007
Hurley, S. (2001). Perception and action: Alternative views. *Synthese 129*, 3–40.
https://pdfs.semanticscholar.org/e147/d24f2f6dc6df2afa20073a977e974e51f186.pdf. Accessed 1/2/2023
Kiverstein, J. (2012). The meaning of embodiment. *Topics in Cognitive Science, 4*(4), 740–758.
Kiverstein, J., Kirchhoff, M. D., & Froese, T. (2022). The problem of meaning: The free energy principle and artificial agency. *Frontiers in Neurorobotics, 16*, 844773. https://doi.org/10.3389/fnbot.2022.844773
Kwisthout, J., Wareham, T., & Van Rooij, I. (2011). Bayesian intractability is not an ailment that approximation can cure. *Cognitive Science, 35*(5), 779–784.
Lanillos et al. (2021). Active Inference in Robotics and Artificial Agents: Survey and Challenges. arXiv:2112.01871.
Legg, C. (2001). Naturalism and wonder: Peirce on the logic of Hume's argument against miracles. *Philosophia, 28*, 297–318.
Linson, A., Clark, A., Ramamoorthy, S., & Friston, K. (2018). The active inference approach to ecological perception: General information dynamics for natural and artificial embodied cognition. *Front Robot AI, 2018*(5), 21. https://doi.org/10.3389/frobt.2018.00021
Lipton, P. (2003). *Inference to the best explanation*. CUP.
Litwin, P., & Miłkowski, M. (2020). Unification by fiat: Arrested development of predictive processing. *Cognitive Science, 44*(7), e12867.
Meloni, M. & Reynolds, J (2021). Thinking embodiment with genetics: Epigenetics and postgenomic biology in embodied cognition and enactivism. *Synthese, 199*, 5415–5416. https://doi.org/10.1007/s11229-020-02748-3
Millidge, B. (2018). Implementing predictive processing and active inference: Preliminary steps and results. https://doi.org/10.31234/osf.io/4hb58
Miller M, Kiverstein J, Rietveld, E. (2020). Embodying addiction: A predictive processing account. *Brain and Cognition, 138*:105495. https://doi.org/10.1016/j.bandc.2019
McCarthy, J. & Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. in b. meltzer & donald michie (Eds.), *machine intelligence 4*(pp. 463–502). Edinburgh University Press.

Morris, D. (2010). Empirical and Phenomenological Studies of Embodied Cognition?. In *The Handbook of Phenomenology and Cognitive Science*, eds. Shaun Gallagher and Daniel Schmicking (Springer Verlag: 2010), 235–252

Nave, K. (2022). Visual experience in the predictive brain is univocal, but indeterminate. *Phenom and Cognitive Science, 21*, 395–419. https://doi.org/10.1007/s11097-021-09747-w

Pezzulo, G., Parr, T., & Friston, K. (2022). The evolution of brain architectures for predictive coding and active inference. *Philosophical Transactions of the Royal Society B*, 377(1844):20200531. https://doi.org/10.1098/rstb.2020.0531

Piersma, T., & van Gils, J. (2011). *The flexible phenotype: A body-centred integration of ecology, physiology, and behaviour*. Oxford University Press.

Rao, R. Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience 2*, 79–87. https://doi.org/10.1038/4580

Raja, V., Valluri, D., Baggs, E., Chemero, A., & Anderson, M. L. (2021). The Markov blanket trick: On the scope of the free energy principle and active inference. *Physics of Life Reviews, 39*, 49–72. https://doi.org/10.1016/j.plrev.2021.09.001

Relihan, M. (2009). Fodor's riddle of abduction. *Philosophy Studies 144,* 313–338.

Sims, A. (2017). The problems with prediction: The dark room problem and the scope dispute. In T. Metzinger, & W. Wiese (Eds.), *Philosophy and Predictive Processing: 23*. MIND Group. https://doi.org/10.15502/9783958573246

Shannon, (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27,* 379–423, 623–656.

Wheeler, M. (2005). *Reconstructing the cognitive world*. MIT Press.

Williams, D. (2018). Predictive processing and the representation wars. *Minds and Machines, 28*(1), 141–172. https://doi.org/10.1007/s11023-017-9441-6