Updating the Frame Problem for AI Research

Lisa Miracchi
University of Pennsylvania

March 18, 2020
Penultimate version: Please cite published version.

> ... the frame problem ... apparently arises from some very widely held and innocuous-*seeming* assumptions about the nature of intelligence, the truth of the most undoctrinaire brand of physicalism, and the conviction that it must be possible to explain how we think. ... One utterly central—if not defining—feature of an intelligent being is that it can "look before it leaps." Better, it can *think* before it leaps. Intelligence is (at least partly) a matter of using well what you know—but for what? For improving the fidelity of your expectations about what is going to happen next, for planning, for considering courses of action ... when we think before we leap, *how do we do it?*
>
> Daniel Dennett, "Cognitive Wheels, the Frame Problem of AI" (1987), p. 44

Introduction

There are many different views about what exactly the Frame Problem is. In computer science and engineering, the Frame Problem is typically understood as that of finding a concise way to represent the effects of actions in a logical system.[1] This problem arose within classical symbolic AI as theorists attempted to find ways of programming artificial agents that could approximate capabilities of humans and other intelligent biological agents. They construed the problem as essentially a logical one: how can we program agents with the knowledge they need to act competently in their environments, so that the correct actions to perform given their programmed goals could be derived by a computer program implementing logical rules? Let's call this the *Narrow Frame Problem* (see Shanahan (2016) for discussion). Several theorists, including some philosophers, have seen in this specific problem a larger question, which deserves to be posed and solved

[1]Thanks to Daniel E. Koditschek, Sonia Roberts, and MIRA Group for their valuable feedback.

at a higher level of generality, and which has important implications for our understanding of all intelligent agency.

We can state this *Generalized Frame Problem* as follows:

> **Generalized Frame Problem.** How can one design a machine to use information so as to behave competently, with respect to the kinds of tasks a genuinely intelligent agent can reliably, effectively perform?

Understood in this way, the Frame Problem is both a deeply theoretical problem and one with clear practical upshots. In the 1980s, it attracted wide attention from philosophers and cognitive scientists, but interest in the problem waned in the '90s along with a loss of interest in genuinely minded (AMI), general (AGI) and human-level (HLAI) artificial intelligence (Russell & Norvig, 2014). A resurgence of interest in these more robust forms of AI – and more generally in the properties of autonomy, robustness, and flexibility that intelligent biological agents exhibit – warrants revisiting the Frame Problem. Moreover, this problem is of interest to cognitive science generally, since it is widely supposed that information processing is crucial for understanding mental processes.

I will argue that the way the Frame Problem is standardly interpreted, and so the strategies considered for attempting to solve it, must be updated. We must replace overly simplistic and reductionist assumptions with more sophisticated and plausible ones. In particular, the standard interpretation assumes that mental processes are *identical to* certain kinds of computational processes, and so solving the Frame Problem is a matter of finding a computational architecture that can effectively *represent* relations of semantic relevance.

Instead, we must take seriously the possibility that the way in which intelligent agents use information is inherently different. Whereas intelligent agents are plausibly genuinely causally sensitive to semantic properties *as such* (to *what* they perceive, desire, believe intend, etc.), computational systems can only be causally sensitive to the formal features that represent these properties. Indeed, it is this very substitution of formal generalizations for genuinely semantic ones that is responsible for the way current AI systems are brittle, inflexible, and highly specialized. Formal causal relationships cannot reproduce the functional properties of genuinely intelligent systems, except in highly specialized and restricted circumstances.

There are two morals we should *not* draw from this lesson. First, we need not abandon the project of building an artificial agent that intelligently makes use of relevant information. I am actually optimistic about

our future ability to solve the Frame Problem, to build artificial agents that reliably and effectively accomplish tasks of the sort genuinely intelligent agents can accomplish.

Second, we need not deny that computational representations play an important role in giving rise to genuinely intelligent information use and behavior. The way intelligent beings use information in solving problems is our *explanandum*. Computational information-processing, on the other hand, is (part of) the *explanans*. What we need is a more sophisticated way of investigating the relationship between the two, so that these two senses of *using information* are not conflated, but instead the question of how they are related to one another can be studied directly.

I approach this problem by applying the *generative methodology* I have developed elsewhere for cognitive science and AI research (Miracchi, 2017, 2019a). According to this methodology, we treat *generative* relationships on close analogy with causal relationships. (A *generates* B just in case B obtains in virtue of A.) This allows us to keep the explanans and explananda in their own vocabulary, specifying them as precisely as possible using *basis* and *agent* models, respectively. Then we can empirically investigate how features of the explanans can make a *generative difference to* features of the explanandum. This allows us to represent complex dependence relations, and so move beyond simplistic assumptions about the relationship between mental and computational states and processes. Applying this approach to the Frame Problem, the question becomes:

> **Updated Frame Problem**: How can features of a system's computational processing, body, and environment be organized in complex and sophisticated ways so that a whole system is generated which can genuinely make use of semantic relevance in acting and planning intelligent actions?[2]

In section 1, I explain why the way genuinely intelligent agents *use information* should be characterized in inherently mental terms that essentially involve intentionality and, frequently, consciousness. In section 2, I review the standard interpretation of the generalized Frame Problem and explain how a problematic conflation of intelligent information use and computational information processing follows from key explanatory commitments of the widely adopted computationalism (Fodor, 1980, 1987; Marr, 1982). In section 3, I explain how the generative methodology can help us to productively re-conceptualize the Frame Problem.

[2] This is a determinate of what I call the "Key Question" in Miracchi (2019a).

1 Intelligent Performance Involves Mental Features

In this section I build on some previous work to explain why AI theorists should conceive of agent information use in an inherently mental way for the purposes of making progress on the Frame Problem. We are looking to understand what is meant by *information use* for a specific kind of task achievement: that characteristic of genuinely intelligent agency. Let's start, then, by getting clearer on what we should mean by "genuinely intelligent agency" and the kinds of tasks that manifest it.

Although many engineers and researchers interested in intelligent agency today are not necessarily interested in building an AI with a *mind of her own* as Ada seems to have in *Ex Machina* (2014) and as Samantha has in *Her* (2013), there are good methodological and objective reasons to give the project of building an artificial minded intelligence (AMI) a special place. First, our epistemic grip on human-level intelligence (and so HLAI) is derivative from our grasp of the intelligent behavior of actual humans, who have minds. We use our assessments of what tasks involve high levels of intelligence when humans accomplish them (such as playing Go or chess), and it is not clear that such assessments of intelligent performance can be divorced from genuinely mental aspects of those performances, such as the phenomenology of effort, creative insight, or intentional strategy.

Plausibly because of this, even when we do succeed in building AI systems that beat humans (e.g. AlphaGo beating Lee Seedol in Go, Silver & Huang (2016)), some critics tend to respond that the AI isn't really intelligent, that the *way* it's playing the game isn't inherently different and more mechanical than the way we do. This suggests that our categorizations of tasks as involving human-level intelligence are complex and difficult to make precise without using mental terminology.

One might think that such rejections reflect mere egotistical or anthropomorphic bias. However, it is not clear why we should trust our intuitions about which tasks indicate intelligence in *choosing* HLAI tasks, but at the same time be skeptical about our intuitions in whether the formalization of performance assessment succeeds in capturing performance of the kind of task we are really interested in. As it currently stands, neither choice of task nor choice of assessment has strong theoretical motivations or constraints.[3] In investigating HLAI and AGI we are epistemically tied to more robust conceptualizations of tasks indicative of *genuinely mental* intelligent

[3]Perhaps if we are successful in designing systematic *agent* models, we can develop some, but it is beyond the scope of this paper to develop the suggestion.

agency whether we like it or not.

While the goal of more rigorously choosing illustrative tasks and assessments is certainly desirable, perhaps our judgments are onto something, reflecting the nuances of a more complex reality than AI researchers have widely acknowledged. I argue in some recent work that genuine mental intentionality and even consciousness cannot be eliminated from our characteristic descriptions of mental processes or behavior without significant loss of accuracy and explanatory power Miracchi (forthcoming b, 2019b). Successes in eliminating inherently mental, content-involving causal generalizations of mental processes in favor of computational, neural, or other non-mental descriptions have been extremely limited (see e.g. Dietrich & List (2016)). Instead, when we investigate contemporary optimism about such reductions, we often find a slip from the highly supported claim that it is widely possible to intervene on mental processes by manipulating lower-level features (such as neural processes, hormones, etc) to, instead, the claim that we have discovered lower-level mechanisms that accomplish mental functions (see Bickle (2015) as an example, discussed in Miracchi (forthcoming b).)

The former claim is not in dispute. What we want to know is whether intelligent task performance and the way intelligent agents use information for such performances can be characterized non-mentally. Not only is there little direct evidence for this, but examples in other higher-level sciences should make us very cautious. Consider thermoregulation, for example.[4] The explanation of *in virtue of what* an organism thermoregulates involves a complex suite of internal mechanisms that are coordinated so that each is activated in the right conditions, and the "result" is that the organism as a whole stays within the optimal temperature range. Thermoregulation has different characteristic causes and effects than its underlying mechanisms, and so can be evolutionarily selected for. Causally intervening at the level of the underlying mechanisms tends not to change whether the organism as a whole thermoregulates, because the underlying mechanisms are carefully orchestrated to compensate for one another.[5] Failure to thermoregulate can cause life-threatening problems, but failures of underlying mechanisms typically do not (another one will step in to accomplish the task.) Selection for thermoregulation will often entail the development of more and more subtly orchestrated lower-level mechanisms, so that the or-

[4]The examples of the Mendelian gene and thermoregulation are ones I have explored in detail elsewhere (Miracchi, 2017, 2019a, forthcoming b).

[5]See also Andersen (2013) for discussion of this issue regarding homeostasis, and it's implications for causal inference.

ganism successfully thermoregulates across increasingly wide conditions.

While thermoregulation depends on a complex suite of mechanisms at lower-level scales, it cannot be identified with them. It is how the organism's suite of mechanisms is sensitively coordinated, in relation to body and environment, that explains the higher-level capacity.

This phenomenon is not specific to biological systems, but is even illustrated with higher-level characterizations of physical systems, such as describing gases using Boyle's Law (Strevens, 2008), for which the dynamics of individual molecules is irrelevant. (Imagine trying to do chemistry purely in the language of physics, or biology purely in the language of chemistry! Good luck!)

What examples like this show is that throughout the special sciences theorists are increasingly accepting that higher-level kinds operate at their own explanatory levels, governed by irreducible causal generalizations. This does not mean (as we'll discuss soon) that we cannot understand how higher-level kinds are generated by lower-level kinds. What it does mean is that, where higher-level kinds play robust causal explanatory roles, understanding relationships between scales may mean that *identifications* will not be forthcoming. Higher-level and lower-level kinds are beholden to different explanatory scales, and as such we should understand resistance to identification not as a failure, but as increased understanding of the differences in explanatory roles and scales between higher-level and lower-level kinds respectively.[6]

The contemporary assumption that we can can carve off the functions and behavior associated with consciousness from their conscious aspects is partially due to David Chalmers (1995)'s distinction between the so-called "easy" problems of consciousness – which are the problems of providing scientific explanations of functions and behavior associated with consciousness – from the so-called "hard" problem, of explaining how non-consciousness-involving processes could give rise to the *what it's like* of experience. As tempting as this may be, I argue in Miracchi (2019b) that this division of problems commits *substitution bias*, unwittingly replacing easier tasks for the ones we want to solve. For example, consider the action of performing a difficult yoga pose, such as *pincha mayurasana*. We have no evidence that such poses can be performed unconsciously. When we explain what we're doing and how we're doing what we do, consciousness seems to be crucially involved. We are keenly aware of whether our weight

[6]See also Griffiths & Stotz (2013) regarding different senses, and corresponding explanatory roles, of "gene".

is more towards the forearms or more towards our fingertips, of how our hips feel in relation to the base that makes contact with the ground. We should not assume at the outset that the kinds of external and internal behavior associated with consciousness can be specified without reference to consciousness.

Consider a purely mental task, like understanding that your ex-spouse is manipulating your children to exert control over you. Drawing that inference relies on conscious reasoning about what the person did, how they acted, what their motives might have been, etc. It also involves (often rather intense) feelings and emotions. When we describe such mental tasks, we do so in ways that are inherently imbued with content and consciousness. This reasoning is not deductive, but rather is inherently content-involving. Moreover, *the way* it is content-involving involves *you*, as conscious thinker, trying to figure out what to believe.

Let us return to the Frame Problem. The lesson I suggest we take from these examples is that if we are interested in building artificial agents with the capacities that are exhibited by genuinely intelligent agents, we should take seriously that (i) our best epistemic handle on the kinds of tasks we're interested in involves intentionality and consciousness, at least in many cases, and (ii) it may be that the performing of the tasks themselves indeed involves all of this *juicy* mental stuff that AI engineers have been dutifully ignoring for the last 80 or more years. Perhaps intelligent agency can't be explained without the kind of desire that phenomenally *pulls* at you to practice (playing Go, or pincha mayurasana), or perhaps the kind of reasoning that involves gut-wrenching concern for your children. Maybe explaining intelligent behavior really is that *hard*.

If so, and we have significantly distorted the Frame Problem by trying to make specification of the kinds of tasks we're interested simpler and more objective by eliminating mental intentionality or consciousness from characterizing them, then we should expect to get stuck in certain systematic ways. This is exactly what we do find, as I'll now show.

2 The Insolubility of the Traditional Frame Problem

One key way in which mental intentionality has been written out of the project of solving the Frame Problem is that theorists have traditionally assumed that content-sensitive descriptions of intelligent processes and behavior are to be replaced by formal descriptions. Daniel Dennett characterizes the project of solving the Frame Problem in this way:

> ... now *how* can ideas be designed so that their effects are what they ought to be, given what they mean? Designing some internal things—an idea, let's call it—so that it behaves *vis-á-vis* its brethren as if it meant *cookie* or *pain* is the only way of endowing that thing with that meaning; it couldn't mean a thing if it didn't have those internal behavioral dispositions. That is the mechanical question the philosophers left to some dimly imagined future researcher.
>
> Dennett (1987), pp. 44-45

The assumption here is that the task of building an AI whose mental states respect semantic relationships (behave the way they "ought" to) is to build an AI with internal symbols whose internal (formal) properties play structurally analogous causal roles ("vis-á-vis its brethren"). The Frame Problem so interpreted asks:

> **Computationalist Frame Problem**: How can we design a system so that its formal properties are isomorphic to its semantic properties (or those we would like it to have), so that those formal properties can account for the system's behavior in the way we would expect if we were taking its semantic properties to be causally efficacious?

This strategy is an instance of the more general strategy of computationalism, according to which mental states and processes are identical to computational states and processes (e.g. Fodor (1980, 1987); Marr (1982), see Clark (2001) for an overview). This means that the kind of agential information use that is our target explanandum for the Frame Problem must be a kind of computational information processing.

The main difficulty with this identification is that mental processes, because they are content-sensitive as we saw in the previous section, are inherently relational. The properties a perception has of being *of a mouse*, or that a belief has of being *about Toni Morrison*, or that an emotion of excitement has of being *for a friend's upcoming visit*, are all relational properties. Semantic characterizations go beyond describing the internal functioning of the system—the agent, or her brain—they relate the agent to her environment. If we must appeal to semantic properties in providing causal descriptions of genuinely intelligent information use, then our causal generalizations are inherently relational.

Computational processes, however, are inherently sollipsistic (Fodor, 1980). A *computational process* is a process that is specifiable entirely in terms of its formal properties. *Formal properties* are either physical properties of a system or abstractions from its physical properties. Formal properties thus are independent from their body- or world- involving relationships. In describing a system, mechanism, or process as computational, we are committing to the claim that we can *carve* off, at least in principle, the role that the environment (including any larger computational environment that might be present) causally plays from it. The causal contributions and effects can be are constrained to be formally specified inputs, and outputs. The body and environment make no contribution to how internal processes are performed.[7]

As others have noted (e.g. Egan (2014)), this important difference between the sollipsism of computational processes and the relationality of their semantic contents is very powerful, and exploitation of this idea is the crucial strategy behind computationalist assumptions in cognitive science. Because computational processes are environment-independent that we can simplify the project of designing computational systems. Thus computational relationships can be, in certain contexts, structurally analogous to semantic relationships, and so useful for accomplishing world-involving tasks. However, the foregoing discussion raises an important concern about this strategy, at least as a methodological assumption, for AI research (and cognitive science generally, for that matter). What if computational processes are more like the lower-level mechanisms underlying thermoregulation, where a system's capacity to thermoregulate is not determined by a single mechanism, but rather by the coordination of internal mechanisms in response to specific bodily and environmental conditions? In such a case there would be no "carving off" the joints of the mental system from the body and environment, and so no prospect of identifying mental processes with computational processes. If mental processes are genuinely content-sensitive, computational processes will be unable to replicate them.

To see the issue more clearly, consider the following semantic generalization that might be used by an intelligent agent: *If a colleague invites you to dinner, you do not show up empty-handed.* The invitation may come in a variety of physical formats: orally, by email, by post (these days only in extraordinary cases), and it may be phrased in any number of ways, typically including more specific content (the inclusion of a partner or children, sug-

[7] Somewhat confusingly, Susan Hurley (1998) refers to this as a kind of *vertical* modularity, and Brooks (1991) as a kind of horizontal modularity.

gestions of dates or menu items). Compliance with the consequent might be instantiated by wine, flowers, dessert, or perhaps an artifact from recent travel or a toy for your colleague's child. For any of these determinates of invitations or gifts, in some other situations they will not count as such. For example, in the context where my colleague and I are also close friends, the very same words may not be intended as an invitation to me, but instead as asking for advice on how to word the invitation to someone else. I might have stopped at the wine store on the way to dinner intending to keep the bottle for myself. I might bring an artifact from recent travel that we had discussed earlier, but not give it as a gift. And so on.[8]

Moreover, although the content-involving rule is counterfactually supporting, there may be no determinate of this rule that supports the same counterfactuals, even in normal conditions. Upon receiving an invitation I may be no more likely to bring flowers than wine, and whether I bring one or the other may have nothing to do with whether the invitation comes by email or conversation, or whether the date is given numerically or by relation to the date of the invitation (e.g. next Thursday). Thus there is no hope of relating determinate e.g. visual or auditory inputs to determinate motor outputs so that a formal algorithm instantiates the computation. So many other factors may come into play in specifying the determinates of invitations and gifts that plausibly the only relevant counterfactual-supporting generalization that can be made is: *if a colleague invites me to dinner, I don't show up empty-handed.*

Formal computations come with a commitment to formally specifiable inputs and outputs that an algorithm can relate, whereas with semantic rules there generally can be no such commitment: the relational contents specifying the rules inherently cross-cut physical, especially proximal, features.

What we lose sight of when we treat semantic generalizations as computational is the substantial work that goes into triggering the right algorithm (or collection of algorithms) in the right situation. The orchestration of algorithms, in coordination with often quite complex bodily and environmental facts, is highly relevant to whether the semantically characterized generalization is instantiated. Because this is so, it is inappropriate to call

[8]It is no accident that computational cognitive science's most specific and convincing examples of content-involving computational rules are ones where the content is mathematical, at the very least, and if there is distal content it is scaffolded onto this mathematically specified content (as in, e.g. Marr & Hildreth (1980)'s model of edge detection to detect zero-crossings and Shadmehr & Wise (2005)'s computational theory of motor control, as discussed in Egan (2014).

such generalizations computational. For the very reasons that generalizations involving mental states and processes need to be semantically characterized, they are not directly amenable to computational identification. Because the right causal description of the process is genuinely semantic, and so genuinely relational, it cannot be instantiated by formal, sollipsistic, processes.[9]

We can now understand why the Frame Problem has been so persistent, and difficult. On the assumption of computationalism, we have ignored the possibility that *agential information use* of the kind crucially involved in effective, reliable intelligent behavior might not be specifiable computationally. We have been trying to *represent* relations of semantic relevance in the hopes that the right representations would enable us to *replicate* agential information use.[10] But if computational processes are inherently causally isolated from relational factors, then they will only be able to functionally mimic inherently relational semantic processes, and they will do so only under very specialized conditions where body and environment are tightly controlled.

When the issue is articulated in this way, we can expect the kinds of failures, and persistent difficulties in solving them, that we have in fact found in AI and robotics research. The massive advances that have been made for more specialized systems or highly controlled environments have not transferred well to these other problems (as e.g. evidenced by difficulties encountered at the DARPA Robotics Challenge, C. G. Atkeson & Xinjilefu (2015)). More complex forms of processing such as deep neural nets have shown progress over classical AI systems, but this technology does not readily transfer to novel problems and on its own and often tends to be slotted into traditional architecture (Marcus, 2018). Increasing the quality of information (e.g. image classification) does not solve the problem. Although there are attempts to use deep neural net technology in robotics to develop robust and flexible behaviors (cf. Ha et al. (2019)), there is to date no general theorizing about deep neural nets that helps us understand how they might solve the problem.

In designing and building such specialized systems we have replaced a

[9] Peacocke (1994) defends the idea that there is a genuinely content-involving conception of computation that cognitive science makes use of, attempting to appeal to normal conditions to produce a parallelism between algorithmic and relational characterizations of mental processes. But as the preceding example shows, appeal to normal conditions is not enough. What we seem to have is the *coordination* of algorithsm to *produce* genuinely relational causal generalizations. Our theorizing should reflect this.

[10] See also Searle (1980).

genuinely relational characterization for a formal characterization, and expect the formal system – whose causal contours are inherently sollipsistic – to behave as if its causal contours were inherently relational. If we really want to build artificial agents that can accomplish tasks that are relationally specified in a wide variety of environments, we should investigate how to build agents with such relational causal profiles directly.

3 Updating the Frame Problem with the Generative Methodology

Let us regroup. I have argued that the way the Frame Problem is standardly interpreted, and thus the strategies considered for attempting to solve it, are overly simplistic and reductionistic. In particular, the standard interpretation assumes that solving the Frame Problem is a matter of finding a computational representational architecture that can effectively *formally represent* relations of semantic relevance. The assumption that one may replace semantic causal generalizations with formal ones and thereby explain intelligent information use fails to account for the inherent relationality of semantic processes, on the one hand, and the inherent sollipsism of formal processes, on the other. Formal representations of semantic relationships can only produce brittle systems that at most reproduce some functional properties of genuinely intelligent systems in highly specialized and restricted circumstances.

I have suggested that instead we should focus our attention on building a system that is genuinely causally sensitive to semantic relevance *itself*. We should focus on building systems whose lower-level parts are orchestrated such that the system as a whole responds to relationships of semantic relevance.

I have argued elsewhere that this kind of project – finding the naturalistic bases of mental kinds – project can be made more precise by understanding it on serious analogy with causal explanation (Miracchi, 2017, 2019a). According to this generative methodology, we treat the empirical investigation of how non-mental, non-intelligent processes give rise to intelligent ones as investigation into a kind of another kind of difference-making relationship. While causal explanations explain how what comes before makes a difference to what comes after, generative explanations explain how that which is more fundamental, or "lower-level" structures itself into what is less fundamental, or "higher-level". These higher-level kinds obtain wholly in virtue of lower-level kinds, but empirical investi-

gation into how this occurs neither requires, nor results in, a definition of higher-level kinds in lower-level terms. Instead, the generative methodology aims for an *a posteriori* understanding of what difference-makers there are, and uses measures of both explanans and explanandum variables in context to infer those relationships.

On this approach, solving the Frame Problem is a matter of finding computational, bodily, and environmental generative difference-makers to how agents use information. Importantly, a generative explanation involves three components. First, there is the *basis model*, the explanation of the computational, bodily, and environmental factors in those terms: their important features and causal roles. It is models of this sort that we are most familiar with in AI and cognitive science. Adequate basis models will plausibly appeal heavily to computational information processing.

Then there is the *agent model*, which describes the intelligent behavior or process we want explained – in our case, use of relevant information for accomplishing a task. Our agent models will aim to describe as clearly and systematically as possible the kinds of intelligent behavior and information use we are interested in artificially replicating, but they will not aim to eliminate mental vocabulary. Adequate agent models will appeal to agents *as users* of information and *performers* of intelligent tasks, and appeal to the agent's understanding and consciousness in describing how the agent uses information in accomplishing her aims.

Lastly, there is a generative model, which describes how manipulating features of the basis model *makes a generative difference to* features of the agent model. Computational information processing is part of the explanans, while intelligent use of information by artificial agents is the explanandum.

The Frame Problem is thus the problem of understanding how computational processing, orchestrated in complex ways with artificial body and environment, can generate genuine semantic sensitivity at the level of the whole agent. We can update the Frame Problem now as follows:

> **Updated Frame Problem**: How can features of a system's computational processing, body, and environment be organized in complex and sophisticated ways so that a whole system is generated which can genuinely make use of semantic relevance in acting and planning intelligent actions?[11]

There are a few things to note right away. First, as a methodology and

[11] This is a determinate of what I call the "Key Question" in Miracchi (2019a).

strategy, there is no guarantee that we will be able to provide a general theory of how intelligent agents use relevant information. We may at most be able to explain how an intelligent agent is able to use relevant information for a certain range of behaviors, in the kinds of contexts she tends to find herself in. At least as a methodological starting point, we should not require more generality that this. Just as in the causal case one can accurately describe causal mechanisms even if they hold only locally (Godrey-Smith, 2008), so in the generative case one might provide a genuine explanation of an agent's use of semantic properties that holds only for a range of circumstances. Another reason to expect this is that the range of basis models will be as diverse as the range of possible AI architectures, bodies, and environments. Perhaps there is something general to say at the physical level, but it may be that all such systems will have in common is that they are organized in ways that give rise to intelligent agent information use.

Second, even though the Updated Frame Problem as specified is inherently more complex than traditional ways of understanding and approaching it, it has the potential to be a rigorous and systematic research program on which incremental progress can be made. Even though we are not currently in a position to tackle anything as difficult or complex as the use of information for behavior that is clearly intelligent, such as performing pincha mayurasana or inferring that your ex is manipulating your children to exert control over you, adopting a generative methodology puts us in increasingly better position to develop more and more sophisticated basis, agent, and generative models. If we increasingly understand how features of computational mechanism relate to body and environment in coordinated ways so as to produce desired behaviors across a range of circumstances, we may eventually be able to understand how to generate the robust and flexible behavior characteristic of genuinely relational information use.[12]

Recently collaborators and I took the first step towards illustrating the viability of this approach. In (Roberts et al., forthcoming), we provide six examples of research in legged robotics that show how a generative analysis can help us to distinguish basis-level commitments (e.g. what representations are needed) from agent-level explananda (in this case Gibsonian affordance-exploitation of the environment, Gibson (1979)). By distinguishing basis and agent models we open up empirical space to ask how

[12]In this way, we can adopt Brooks (1985, 1991)'s insight that AI research will be more productive by incrementally developing systems with real-world capabilities, instead of designing systems that are supposed to play a role in an armchair-specified architecture. This insight does not require his accepting his general anti-representationalist stance.

information-processing might generate affordance exploitation, instead of supposing that the desired behavior must be, or is most effectively produced by, certain kinds of representational states. We show how capacities for simpler behaviors can be organized in order to generate systems characterizable as exploiting relational affordances, but these affordances are not explicitly represented by the robot. We also show how advances can be progressively built upon to produce robots with more complex and robust capacities (see, e.g. Ilhan et al. (2018)).[13]

The prospects of this kind of empirically based incremental approach – that includes explicit philosophical and empirical attention to specifying the kinds of tasks of interest and aims to specify ever more precise and accurate agent models, thereby opening up the space for empirical investigation into generative relationships – are bright. The Frame Problem might be hard, but it no longer seems hopeless.

4 Conclusion

I hope to have shown why the Frame Problem has seemed so intractable, why oversimplifying and reductive assumptions are the root of the difficulty, and how to update those assumptions in a way that helps us approach the Frame Problem in a new, more productive light. The Frame Problem can be solved, just not as traditionally conceived. Instead of assuming that agential information use can be given a computational specification directly, we should look to provide generative models of how computational information processing, in body and environment, generate this relational capacity that we and other genuinely intelligent beings have.

I close with brief discussion of an issue that I have before now left implicit, namely how the preceding discussion gives robotics a central place in AI research. Because semantic generalizations are relational, a system that is inherently semantically characterizable cannot have the kinds of causal joints that disembodied AI systems have. Indeed, the only empirically tractable way I see of generating a system that is inherently semantically characterizable is by starting with simpler robotic systems and progressively making them perform activities that increasingly approximate intelligent behavior. Real progress on the Frame Problem will be made, not by developing systems that can beat us at specific human-level tasks, but

[13]The last two examples use explicit representation of affordances at the most abstract level, but the effectiveness of these representations crucially depends on lower-level affordance-exploitation that does not involve explicit representation of their corresponding affordances.

by investigating how artificial embodied systems might be progressively designed so that they increasingly interact with their environments as if they have minds of their own, using what is available to them to get their aims accomplished.

References

Andersen, H. (2013). When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, *80*, 672–673.

Bickle, J. (2015). Marr and reductionism. *Topics in Cognitive Science*, *7*, 299–311.

Brooks, R. A. (1985). A robust multi-layered control system for a mobile robot. Ai memo 864, Massachusetts Institute of Technology.

Brooks, R. A. (1991). Intelligence without representation. *Artificial Intelligence*, *47*(139–159).

by Alex Garland, D. (2014). *Ex Machina*. A24 Films.

by Spike Jonze, D. (2013). *Her*. Warner Brothers.

C. G. Atkeson, B. P. W. Babu, N. B. D. B. C. P. B. X. C. M. D. R. D. S. F. P. F. M. G. J. P. G. P. H. A. J. K. K. J. K. L. L. X. L. C. L. T. P. F. P. G. G. T. & Xinjilefu, X. (2015). What happened at the darpa robotics challenge, and why.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219.

Clark, A. (2001). *Mindware*. Oxford University Press.

Dennett, D. (1987). Cognitive wheels: The frame problem of ai. In Z. Pylyshyn (Ed.), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence*. Ablex.

Dietrich, F. & List, C. (2016). Mentalism versus behaviorism in economics: A philosophy-of-science perspective. *Economics and Philosophy*, *32*, 249–281.

Egan, F. (2014). How to think about mental content. *Philosophical Studies*, *170*(1), 115–135.

Fodor, J. (1980). Methodological solipsism as a research strategy in cognitive science. In R. Boyd, P. Gasper, & J. D. Trout (Eds.), *The Philosophy of Science (1991).* (pp. 651–669). MIT Press.

Fodor, J. (1987). *Psychosemantics*. MIT Press.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Houghton-Mifflin.

Godrey-Smith, P. (2008). Reduction in real life. In J. Howhy & J. Kallestrup (Eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*. Oxford University Press.

Griffiths, P. & Stotz, K. (2013). *Genetics and Philosophy: An Introduction*. Cambridge University Press.

Ha, S., Xu, P., Tan, Z., Levine, S., & Tan, J. (2019). Learning to walk in the real world with minimal human effort.

Hurley, S. (1998). Alternative views of perception and action. In *Consciousness in Action* chapter 10. Harvard University Press.

Ilhan, B. D., Johnson, A. M., & Koditschek, D. E. (2018). Autonomous legged hill ascent. *Journal of Field Robotics*, *35*, 802–832.

Marcus, G. (2018). Deep learning: A critical appraisal. *ArXiv:1801.00631, [Cs, Stat]*.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W.H. Freeman and Company.

Marr, D. & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society*, *207*, 187–217.

Miracchi, L. (2017). Generative explanation in cognitive science and the hard problem of consciousness. *Philosophical Perspectives*.

Miracchi, L. (2019a). A competence framework for artificial intelligence research. *Philosophical Psychology*, *32*(5), 589–634.

Miracchi, L. (2019b). None of these problems are that 'hard' ... or 'easy': Making progress on the problems of consciousness. *Journal of Consciousness Studies*, *26*(9-10), 160–72.

Miracchi, L. (forthcoming-b). Embodied cognition and the causal roles of the mental. In ed. Michael Brent (Ed.), *Mental Action*. Routledge.

Peacocke, C. (1994). Content, computation, and externalism. *Mind and Language*, *9*(3), 303–335.

Roberts, S. F., Koditschek, D. E., & Miracchi, L. (forthcoming). Examples of gibsonian affordances in legged robotics research using an empirical, generative framework. *Frontiers in Neurorobotics*.

Russell, S. & Norvig, P. (2014). *Artificial Intelligence: A Modern Approach* (3rd ed.). Prentice-Hall.

Searle, J. (1980). Minds, brains and programs. *The Behavioral and Brain Sciences*, *3*, 417–57.

Shadmehr, R. & Wise, S. (2005). *The computational neurobiology of reaching and pointing: A foundation for motor learning*. MIT Press.

Shanahan, M. (2016). The frame problem. *Stanford Encyclopedia of Philosophy*.

Silver, D. & Huang, A. e. a. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Strevens, M. (2008). *Depth*. Harvard University Press.