Ergo AN OPEN ACCESS JOURNAL OF PHILOSOPHY

# A Forward-Looking Theory of Content

CAMERON BUCKNER
*Department of Philosophy*
*The University of Houston*

In this essay, I provide a forward-looking naturalized theory of mental content designed to accommodate predictive processing approaches to the mind, which are growing in popularity in philosophy and cognitive science. The view is introduced by relating it to one of the most popular backward-looking teleosemantic theories of mental content, Fred Dretske's informational teleosemantics. It is argued that such backward-looking views (which locate the grounds of mental content in the agent's evolutionary or learning history) face a persistent tension between ascribing determinate contents and allowing for the possibility of misrepresentation. A way to address this tension is proposed by grounding content attributions in the agent's own ability to detect when it has represented the world incorrectly through the assessment of prediction errors—which in turn allows the organism to more successfully represent those contents in the future. This opens up space for misrepresentation, but that space is constrained by the forward-directed epistemic capacities that the agent uses to evaluate and shape its own representational strategies. The payoff of the theory is illustrated by showing how it can be applied to interpretive disagreements over content ascriptions amongst scientists in comparative psychology and ethology. This theory thus provides a framework in which to make content attributions to representations posited by an exciting new family of predictive approaches to cognition, and in so doing addresses persistent tensions with the previous generation of naturalized theories of content.

## 1. Introduction: Naturalizing Mental Content

According to a venerable story, rational agents act on the basis of what they desire and how they believe they can obtain it, and many of their actions can be explained by showing how they are appropriate given the contents of these mental states. A theory of mental content tries to clarify one component of this nexus by specifying how mental states like beliefs and desires come to be about

**Contact:** Cameron Buckner <cjbuckner@uh.edu>

https://doi.org/10.3998/ergo.2238

particular aspects of the world. While some theorists ground content in other intentional entities like consciousness or third-person interpreters, content naturalists attempt to ground it in non-intentional phenomena studied by scientific disciplines like psychology, neuroscience, or evolutionary biology. The attempt to naturalize mental content in this way enjoyed an enormous amount of attention in philosophy of mind from 1980–1995, which I will hereafter refer to as the "heyday" of naturalism about mental content.

One family of heyday views which has retained significant popularity until the present time is called "teleosemantics" (Shea 2013). Teleosemantics is a "backward-looking" naturalistic approach that attempts to ground content in causal or informational relations that held at some point in the agent's past, during its evolutionary or learning history. Most versions attempt to identify a process of selection such as evolution or learning that bestows contents on neural states based on their ability to signify or indicate aspects of the organism's environment that satisfy biological or psychological functions. Opinions today differ on the degree of success achieved by heyday efforts in teleosemantics (the most influential views being those of Dretske 1981; 1988; Millikan 1984; 1989; Papineau 1988).[1] Though some now think heyday efforts presented us with an embarrassment of theoretical riches (Neander 2017; Shea 2013; 2018), others hold that these views failed to achieve the grand ambitions under which they were offered (Chemero 2011; Hutto & Myin 2012; Mendelovici & Bourget 2014; Ramsey 2007). My goal here is neither to endorse nor rebut these criticisms, but rather to defend a newer style of theory which is closely related to teleosemantics, and which differs from the backward-looking heyday views in ways relevant to the most influential objections against them.

The most common concern alleges that teleosemantics faces insoluble indeterminacy (or "disjunction") problems, the worry being that there are too many candidate contents which could be derived from an organism's history, and there is no principled way to select amongst them (Fodor 1987; Summerfield & Manfredi 1998). In the interests of efficiency, it will be useful to focus on one of the most popular views which is closely related to the novel view presented here: Dretske's informational teleosemantics (Dretske 1988). Dretske's theory crucially relies on a distinction between triggering causes and structuring causes in the production of behavior. On this view, reinforcement learning recruits indicators—neural states which stand in informational relations with environmental properties—to control behaviors on the basis of contingencies between the activation of those states and desired outcomes in the environment. Dretske

[1] On most versions of Millikan's view, a representation's contents are derived from the proper function of the systems that consume the representation. It is an interesting question whether Millikan's view could allow consumer systems to bestow forward-looking contents, but Millikan's views on this topic are complex and there is not space to explore this possibility here.

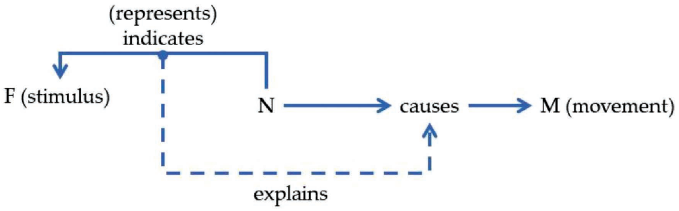

**Figure 1.** Dretske's structuring cause solution (Dretske 1988: 88).

uses the recruitment process to fix a representation's content by supposing that the recruitment process invests the neural state (the triggering cause) with the function of representing the state of affairs that its ability to indicate caused it to be recruited. Since reinforcement learning is a process which is plausibly sensitive to such environmental contingencies, these informational relations provide "structuring cause" explanations for behaviors which are later triggered by these neural states. This solution is ingenious; it purported to both determinately fix a representation's content and render those contents causally relevant to the explanation of later behaviors, all while drawing only upon the naturalistic resources of information theory and reinforcement conditioning (see Fig. 1).

Unfortunately, this view has difficulty simultaneously assigning determinate contents and allowing for a robust possibility of misrepresentation. The difficulty can be highlighted by focusing on ambiguity throughout Dretske's oeuvre on the crucial notion of indication—specifically, between a strict notion of indication that would require perfect correlation during recruitment, and a weaker notion that would allow for something less (Godfrey-Smith 1992). This ambiguity can be used to develop a dilemma argument against the view. On the first horn, the strict interpretation of indication allows it to offer a strong response to the problem of indeterminacy, but only at the cost of being unable to offer an empirically plausible account of misrepresentation. On the other horn, the weaker notion of indication grounds a more plausible approach to misrepresentation, but only at the cost of being unable to offer a strong solution to the problem of indeterminacy. Let us consider each horn in turn.

The earlier versions of Dretske's account of indication required perfect (perhaps even nomologically necessary) correlation between an indicator (*N*) and the state of affairs it indicates (*F*)—in other words, $P(F|N) = 1$ (Dretske 1981). This version of the view can offer a strong response to the charge of indeterminacy: a representation indicates whatever it was perfectly correlated with during recruitment (for a number of worked examples, see Prinz 2000). This view does attempt to offer an account of misrepresentation: here, a representation misrepresents when it is tokened by something other than its content in an environment other than the one in which it was recruited (because the representation must be taken out of the environment in which $F$ is never activated without $N$ for such a circumstance to occur).

Unfortunately, this account of misrepresentation rests on two unrealistic assumptions about the nature of learning. First, the tokening of a representation is almost never perfectly correlated with its content during learning; agents frequently make mistakes during training, learning can satisfice with lower degrees of correlation on the basis of cost/benefit analyses, and noisy, dynamic, and deceptive environments rarely present organisms with perfectly reliable signals for interesting contents in the first place (Godfrey-Smith 1992; Slater 1994). Second, studies of expertise show us that learning is a lifelong continuous process, with agents able to continually improve the causal contact between representations and their contents over the course of decades (Ericsson & Smith 1991; Fodor 1987). Dretske was aware of the difficulties this earlier version of the view had with a psychologically-plausible account of misrepresentation (Dretske 1986), and so may be seen to base his later views on a weaker notion of indication and a more flexible notion of recruitment.

Let us then consider the other horn of the dilemma, explored by Dretske in this later work, where indication requires something less than perfect correlation (e.g., $P(F|N)<1$). Dretske does not define this weaker notion of indication as clearly as he did the earlier, stricter notion (Godfrey-Smith 1992), but it is clear that this version of the view leans harder on the notion of a representational function to settle content indeterminacy. The weaker notion of indication offers more flexibility to accommodate misrepresentation; for example, even during learning, things other than the state of affairs the indication of which caused recruitment might token the representation. Unfortunately, the same neural state will typically stand in weaker, overlapping indication relations to many different environmental properties simultaneously, and the view can no longer lean on perfect indication to arbitrate which of these weaker correlations picks out the content that caused it to be recruited (Fodor 1990; 1994; Summerfield & Manfredi 1998). In particular, we may worry about choices between ascribing more modest, appearance-like contents and more ambitious contents that seem more closely-related to the benefits obtained by the organism during recruitment that led to reinforcement (Neander 2006). For example, suppose a representation which controls a bird's consummatory movements simultaneously indicated (to differing degrees) "housefly", "fly", "prey", and "small, dark, moving speck" during recruitment. Reinforcement learning is sensitive only to the degree of correlation between environmental stimuli and rewards or punishments, and each of these properties stands in some degree of contingency with the relevant rewards. Thus, even in what is often regarded as the most promising heyday effort, the problems of indeterminacy and misrepresentation are locked in a tension that is difficult to resolve.[2]

[2] In some ways, failing on indeterminacy is worse than merely offering an implausible account of misrepresentation, for indeterminacy also undermines any account of misrepresentation. As

Despite (or perhaps because of) these tensions, teleosemantics is due for updates since the heyday. Its philosophical motivations remain compelling, and in the interim much has changed in our broader theorizing about the computational architecture of the mind, especially regarding the nature of the learning processes which might shape representations and recruit them to positions of behavioral control. A new generation of philosophers have begun to challenge old assumptions, exploring new territory in naturalized psychosemantics generally and in teleosemantics specifically (Eliasmith 2005; Gładziejewski 2016; Grush 2004; Morgan 2014; Nanay 2014; Rupert 1999; Ryder 2004; Shea 2007a; 2018; Usher 2001). Particularly relevant to the view to be presented here, significant progress has been made regarding predictive approaches to learning and self-supervised neural-network based approaches to artificial intelligence. A debate has taken place as to whether these predictive approaches should be regarded as representational or not; but setting that aside, few philosophers have attempted to update learning-based teleosemantics to fit with predictive approaches to the mind (though see Gładziejewski 2016; Williams 2018).

This paper is hardly the first to offer a naturalist theory that emphasizes the importance of error-correction mechanisms to representational theorizing. Notable precursors include Bickhard (1993), Ryder (2004), and more recently Bielecka and Miłkowski, who have argued that the epistemic evaluation performed by error-correction and anticipatory mechanisms grounds a distinctive causal role for representations to play in cognitive systems (thus emphasizing responses to objections like Ramsey's Job Description Challenge and Hutto & Myin's Hard Problem of Content—see Miłkowski 2015; Bielecka & Miłkowski 2019). Here, I focus more on the problems of misrepresentation and indeterminacy, arguing that by grounding contents in prediction-error learning—which involves top-down prediction of expected inputs (primarily derived from perception, but perhaps also by interoception and the result of internally-simulated inputs as in imagination), comparison of those expectations with actual inputs, and revision of representational structure in response to prediction failure (to better represent the world going forward)—a theory can be offered which boasts distinctive strengths where backward-looking views like Dretske's have been thought to run into trouble.

I call the family of theories inspired by this insight "forward-looking"; their shared core is an emphasis not on informational relations chosen from some idealized past or counterfactual present, but rather from the likeliest stable future. Specifically, the forward-looking approach grounds representational content in the agent's own ability to detect when it has represented the world

Dretske himself notes, "without a determinate [representational] function, one can, as it were, always exonerate [a representational system] of error, and thus eliminate the occurrence of misrepresentation by changing what it is supposed to be indicating" (Dretske 1988: 69).

incorrectly—which in turn allows the organism to more successfully track aspects of that world in the future. Briefly, a representation's forward-looking content is thus whatever a predictive learning system reliably tends to put it in better informational contact with over time in response to prediction error.

I motivate and explain this view in six steps. Section 2 situates the view with respect to some major distinctions in theories of content that have been drawn in the literature since the heyday. Section 3 provides an empirically-grounded inspiration for the view by drawing upon debates over content ascriptions in cognitive ethology and comparative psychology, focusing especially on learning processes more flexible and complex than simple instrumental conditioning. This is significant, since the view will focus on assigning contents to whole agents (or at the "personal" level)—as is common practice in comparative psychology and ethology—rather than focusing more on subpersonal systems as may be more popular in other areas of computational psychology and has been the focus of other recent updates to teleosemantics (e.g., Neander 2017). Section 4 then sketches the forward-looking view as a way to align solutions to the problems of indeterminacy and misrepresentation by grounding content ascriptions in the agent's own capacities to detect when it has misrepresented. Section 5 illustrates the consequences of the view by showing how it handles some familiar and novel test cases. Section 6 concludes by explaining the causal import of forward-looking content to the explanation of behavior, by tying it to some classic discussions by James and showing how it can provide inspiration for experimental designs that help arbitrate active disputes about content ascriptions in comparative psychology and ethology.

## 2. Situating the View in the Literature

I begin to introduce the view and its aspirations by situating it with respect to four distinctions that have been drawn between different approaches to content since the heyday. First, there is a distinction between theories vindicating contents ascribed by folk psychology and those vindicating contents ascribed by empirical psychology. Second, there is a distinction between contents ascribed to representations at the level of the whole agent and contents ascribed to representations at the level of the organism's cognitive or computational parts. Third, there is a distinction between views that ascribe ambitious contents that leave wide room for error, and views which ascribe more modest, appearance-like contents which are rarely tokened in error. Fourth, a distinction has been drawn between monolithic theories that argue for a single kind of content and more multifarious theories of representation that ascribe different tiers of content to different types of selection or learning. I will say a bit about each distinction briefly in turn.

First, heyday views (especially Fodor 1987) often set out to vindicate folk psychology, and many of the key examples from older debates focused on examples drawn from folk ascriptions (Rupert 2018). This approach casts the effort in a more "philosophy of mind" bent, and it diverted much of the heyday discussion to focus on the nature and plausibility of folk psychology itself (especially propositional attitude ascriptions), such as whether folk psychology has the status of a scientific theory and whether it would ultimately be eliminated in the face of scientific discoveries by neuroscience (Churchland 1981; Ramsey, Stich, & Garon 1990). By contrast, Cummins (1990; Cummins, Putnam, & Block 1996) suggested that we should more profitably adopt a "philosophy of science" attitude towards these issues by treating 'representation' as a theoretical posit in cognitive science, and theorizing about the nature of representations and representational content only on the basis of what is required to explain the success of empirical cognitive science. I will be adopting this latter attitude here; however, I will focus on the contents ascribed to flexible learning processes in the fields of comparative psychology and cognitive ethology, where successful empirical work has been explicitly designed around the scientific redeployment of our folk ascriptive capacities on non-human animals (Dennett 1983; Ristau 1991; Timberlake 2007).

Second, others have drawn a related distinction between ascribing contents at the level of whole agents and ascribing contents to an agent's computational parts (Rowlands 1997). Rowlands in particular proposes that many of the indeterminacy problems attributed to teleosemantics can be located instead in a mismatch between proper functions of a cognitive system's components (such as edge-detecting cells in visual areas V1), and proper functions that are better ascribed to whole animals (such as locating environmental affordances that promote survival and reproduction). Here, although I focus on neural states as representational vehicles, the contents ascribed to those vehicles will be attributed at the level of whole agents. Though predictive learning may also operate independently at the level of individual subsystems (such as motor representations), I focus here on whole-agent predictive learning that involves ongoing interaction between central learning processes and the outcomes of predictions (which will typically be determined by the agent's engaged exploration of its environment, but may also include interoception or episodes in internal mental life such as reflection, remembrance, and imagination—Sims 2017). The representational functions assigned to vehicles at this level invoke the reduction of global, whole-agent prediction-error, which as we will see implicates the cooperation of a potentially open-ended number of the agent's other subsystems and interaction with numerous environmental contingencies.[3]

[3] Two other views which ascribe representations to prediction error-correction models are worth mention here: Rupert's (2011) claim that forward models of motor movement are representational, and Shea's (2012) claim that the representations of reward in predictive reinforcement learning models are metarepresentational. However, these views both focus on ascribing contents to subpersonal computational components of the predictive learning system, rather than to the

Third, when faced with pressures over content indeterminacy, some teleosemantic views tend to favor more modest, appearance-like contents, whereas others reach for more ambitious contents that allow for a greater degree of error (Neander 2006). A curious thing about representational hypotheses in cognitive science is that we do not automatically regard a representational ascription as disconfirmed when it fails to predict behavior. For example, my APPLE concept may remain aimed at *apples* even if I occasionally apply it to non-*apples*, such as *wax fruit* or *roundish pear*.[4] Compare this to a hypothesis in the physical sciences like "electrons are negatively-charged"; any case where an electron fails to attract a positively charged particle in the relevant way would be treated as prima facie disconfirmation of the theory, rather than as the electron having made a "mistake". Some views react to this curious aspect of representational hypotheses by ascribing ambitious contents (like *apple*) that explain successful action but outstrip the agent's actual discriminative abilities, while others ascribe instead more modest, appearance-like contents (such as *apple-looking*) that are more closely tethered to what the agent can reliably discriminate. The modest ascriptions look most empirically promising when we are considering simple components like edge-detecting cells in early vision or organisms with very inflexible learning abilities like frogs, but the more ambitious ascriptions seem more appropriate when we move to more sophisticated organisms that have better ways of detecting and responding to representational error.[5] I here focus on ascribing more ambitious contents. However, a challenge (which will be addressed below) is that the more ambitious contents seem to take us further from the agent's current causal organization, from the discriminations it can reliably make in the environment it now occupies. In so doing, indeterminacy problems can be rendered more severe.

Fourth, the view can be situated with respect to an increasingly popular multipronged or pluralist attitude toward content, which locates different "tiers" of content in different representational systems of differing degrees of flexibility and sophistication. Several recent thinkers have recommended this attitude as a solution to post-heyday malaise: that is, they hold that there are different kinds of content that are attributed in different scientific contexts, and there is probably no need to look for the single criterion that unites them all (Godfrey-Smith 2004; Shea 2013). Perhaps there is a kind of basic "Tier 1" biological content that arises

whole organism, as the view proposed here does. The present view is thus closer to original heyday ambitions.

[4] I follow conventions in this literature and use smallcaps to denote representational vehicles (specifically, their types), and italics to denote the properties to which those vehicle types refer.

[5] Another take on modest ascriptions here has them being more action-oriented rather than appearance-like; this approach has been considered especially relevant to predictive processing, though I do not pursue it here. For a review of such recent views, see Williams (2018).

from natural selection and supports adaptive but relatively inflexible behavior as is found in innate-releasing mechanisms studied in ethology (Brigandt 2005); another deflationary, indicator-style "Tier 2" content arising from the kind of simple forms of associative learning described by Dretske (1988); and then at least one more flexible Tier 3 of content that arises from yet more sophisticated epistemic regulation like prediction-error learning; and perhaps even a later Tier 4 content that comes after the addition of social cognition and language (Godfrey-Smith 1992). The positive view offered in this paper can be understood as a way to build the base case of Tier 3 representation. Interestingly, this Tier 3 content might be bootstrapped from Tier 1 and 2 contents by comparing the expectations they generate to incoming sensory information streaming in from the world, as we will see later on.

## 3. A Test Case for Tier-3 Contents: Whole-Organism Attributions in Cognitive Ethology and Comparative Psychology

To guide the development of the theory, I propose a fresh, philosophy-of-science-based test case. In some ways, Dretske's emphasis on the most basic forms of reinforcement conditioning was odd from the beginning, given that comparative psychologists rarely describe behaviors driven by simple conditioning in representational terms. Instead, they tend to reserve the representational idiom for causes of behavior that are more flexible and "insightful" (Buckner 2011). In fact, research progress in these areas frequently breaks down over interpretive debates akin to the heyday disjunction challenges, with some research groups favoring more modest and others favoring more ambitious contents in their ascriptions. Ideally, a naturalistic theory of content should help scientists resolve these debates by providing inspirations for new experiments that could confirm or disconfirm the various ascriptive hypotheses on offer.

A prominent example of such an interpretive dispute concerns the landmark experiment of Hare, Call, Tomasello, and Agnetta (2000) on Theory of Mind in chimpanzees, which has by now been locked in interpretive stalemate for two decades (Heyes 2015; Lurz 2011). In this experiment (Fig. 2), a subordinate and a dominant chimpanzee sit across a shared enclosure from one another, held in their respective rooms by guillotine doors which can be raised and lowered by the experimenters. Food objects can be placed either out in the open (easily visible to both competitors) or behind an opaque occluder from the perspective of the dominant. The guillotine doors are then raised (with the subordinate given a small head start), and experimenters see whether the subordinate reliably takes food objects behind the opaque occluders (which the dominant cannot see) but

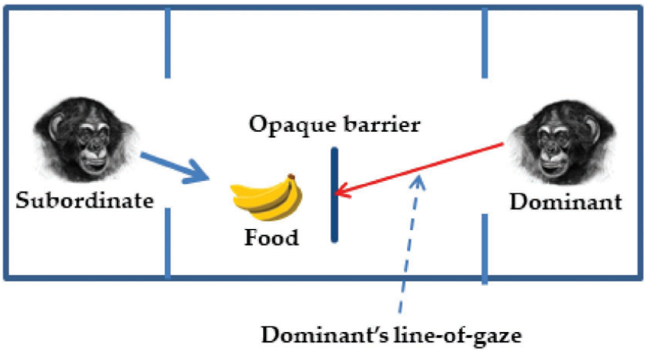

**Figure 2.** The design of Hare et al. (2001).

avoids food objects out in the open (which can be seen by the dominant and thus would become the subject of a dispute if taken, earning the subordinate a beating). Across an exhaustive set of control conditions, subordinates have shown themselves more likely to take food objects that dominants cannot see.

These findings have been interpreted differently by different camps. "Proponents" of animal Theory of Mind argue that they show that the subordinates possess a basic form of perception-goal psychology, representing what the dominant chimpanzee does and does not *see (Call & Tomasello 2008)*. An equally distinguished list of "skeptics", by contrast, hold that such experiments cannot even in principle show that animals represent *seeing*, for the findings can be as well explained with a more modest representation of the overt behavioral cue, *line-of-gaze*—that is, a spatial line beginning at the surface of an animal's head or eyes and terminating in an opaque object (Penn & Povinelli 2007). Dozens of experiments on a half dozen different species—together with numerous subtle essays on experimental design, parsimony, background evidence, and statistical methodology (see Lurz 2011 for a review)—all failed to put this controversy to rest. Moreover, the different camps have not even been able to agree on a hypothetical experimental design that that could arbitrate between these competing hypotheses—and many experiments proposed by skeptics to overcome the deadlock have been ruled insufficient by other skeptics before they were ever even performed. That different camps continue to agree on all the data but disagree on their interpretation suggests that this dispute is not wholly empirical, but rather reflects a deeper (and mostly implicit) disagreement over psychosemantics (Buckner 2014).

Where such interpretive impasse stubbornly resists experimental resolution, science may benefit from philosophical reflection as to what behavioral or neural data ought to even count as evidence for one or another content attribution—and, further, why. This impasse thus offers a new mandate for content naturalism, one which would at least rule out forms of evidence that are not assessable using controlled and reliable empirical methods. Ideally, the new criterion should also come with a defense which explains why cognitive science would be more empirically successful or fruitful if content is reduced in this way, rather than in others—for scientists tend to be a practical lot, and are unlikely to abandon their implicit theories of content on the basis of intuitions or a priori argument alone. The "forward-looking" theory described here is offered as just such an ecumenical psychosemantics with a clear pragmatic defense, which will be provided after the view is introduced in the next section.

## 4. Forward-Looking Content

### *4.1. Background: What is the general shape of a forward-looking theory?*

The key insight of a forward-looking theory of content is that content attributions can be grounded not just in the agent's past or current discriminative abilities, but rather also in the epistemic capacities appropriate to evaluating those discriminations, and especially in how those evaluations will stabilize over time. Such views are immensely powerful, for they open up significant space for misrepresentation—they can allow for a wide gap between past or current discriminatory abilities and contents—but they constrain that gap with forward-directed epistemic abilities that behavioral experiments could confirm actually operate on the representations in question.[6] In short, organisms need not be so perfect

[6] This essay is not the first to offer a forward-looking theory of content—and the view offered here is inspired by the germ of a forward-looking view offered by Dretske himself (Dretske 1986)—but the most developed options of which I am aware thus far are non-naturalist. Consider the core principle of Dickie's (2015) view:

> Principle connecting aboutness and justification (approximate version) S's <$\alpha$ is $\Phi$> beliefs are about an object iff their means of justification converges on the object, so that, given how the beliefs are justified, the subject will be unlucky if they do not match the object and not merely lucky if they do.

On the central issues of this essay, Dickie's view serves as a break from traditional causal or descriptive theories of content; mental reference on her view is grounded not in any features of a representation's causal history or current informational relations, but rather in the means of justification appropriate to evaluating a belief and what they would tend to non-accidentally confirm over time.

To provide another example, a different kind of forward-looking theory can be found in the work of Mendelovici, who grounds content attributions not in notions of "luck" and "justification", but rather in what the owner of that state takes herself to mean through "cashing out" intuitions that are accessed only after relevant alternatives are consciously considered:

> Taking: Subject S takes immediate content C to mean C+ if S is disposed to have a set of cashing out thoughts that together specify that C cashes out into C+ (Mendelovici 2018: Ch. 7)

as to never make mistakes—not even in an idealized learning history—but what counts as a mistake is determined not by the theorist's idealizations of the past, but rather by the agent's own justificatory or metacognitive abilities. Such views can moreover be regarded as naturalist if we can locate a general mechanism that psychology and neuroscience tells us implements the relevant forms of forward-directed epistemic processing.

Let us thus consider the empirical and philosophical work on such epistemic mechanisms. The best candidates are those that support various forms of error-correction learning, which grant agents the kind of cognitive flexibility that ethologists and comparative psychologists think merit mentalistic description (Buckner 2015). Organisms that can detect representational failures have access to a form of evidence that allows them to continually improve the causal contact between a representation and its referent, and empirical evidence of such "self-monitoring"—from learning curves to looking time data in "dishabituation paradigms"—can be used by cognitive scientists as operational evidence of a representation's content (Allen 1999). Most artificial models of error-correction learning in cognitive science are "supervised", where an external signal tells a model whether it got the right answer. However, we should not rely on supervised models here, for they require a programmer or oracle to perform the very semantic work that needs explaining—that is, how agents themselves could possess contents that outstrip their present discriminative abilities. Only by uncovering mechanisms that allow organisms to compute their own prediction errors can we finally locate the causal/explanatory role played by misrepresentation itself.

Such mechanisms can be found in a popular wave of recent work in neuroscience and psychology, especially in predictive coding models in theoretical neuroscience and predictive (or "self-supervised") deep-learning networks in artificial intelligence. Several influential philosophers of cognitive science have been impressed by the potential of such models (Clark 2015; Hohwy 2013; Ryder 2004). These views take a variety of specific forms—some are abstractly mathematical (Friston 2010), others more mechanistic and neuroanatomically located (Gluck & Myers 2001; Ryder 2004); some are hierarchical and involve many layers of processing (Luc, Neverova, Couprie, Verbeek, & LeCun 2017), whereas others involve only a single layer of cortical neurons (Favorov & Ryder 2004). In the interests of space—and since numerous philosophically-informed reviews from a wide variety of theoretical perspectives are readily available (Clark 2015; Hohwy 2013; Seth 2014; Williams 2018)—I will not survey the details of these views here. The important shared core for the forward-looking theory offered

In other words, on Mendelovici's view whether a representation has one content or another is determined by which contents its owner is disposed to accept after considering the alternatives and reflectively stabilizing on a set of self-interpretive judgments.

here is that the overarching "goal" of such processing is not just to infer the correct action from "bottom-up" processing on incoming inputs alone (which, again, will primarily come from the senses but may also be generated by internal systems such as interoception, memory, or imagination), but rather also to correctly predict and control those incoming inputs using "top-down" processing guided by prediction failures.[7] This kind of learning thus offers a dramatic shift away from the kinds of basic evolutionary or reinforcement learning mechanisms emphasized by heyday teleosemantic views, and it is only this shared predictive core that I will draw upon below.[8]

## 4.2. The Theory

I now develop and explain the preferred forward-looking theory of content and representation.[9] In doing so, it will help to anchor the discussion to a particular

[7] A reviewer notes that this distinction may call to mind another drawn between model-free and model-based approaches to learning and planning in cognitive science and machine learning (e.g., see Sutton & Barto 2018: Ch. 7 and 14). I shy away from that terminology here, however, as it is not entirely clear at present whether model-based learning requires some dedicated component of the network to explicitly represent a model of the world's future states (and what, exactly, "explicit" might mean here), or whether model-based learning could be implemented implicitly by simpler mechanisms with sufficient depth or scale (e.g., as in some deep reinforcement learning neural network models that have recently pushed the boundary between model-free and model-based approaches—Botvinick et al., 2019; Gershman, 2018; Kulkarni, Saeedi, Gautam, & Gershman 2016). This uncertainty may also be present in some of the more ambitious speculations about the power of simple associative learning mechanisms at scale, e.g., as expressed by Dretske in the later chapters of *Explaining Behavior*.

[8] An initial complication is that researchers studying predictive learning disagree as to whether this work supports a representational or anti-representational interpretation. I will not enter directly into this debate here; for defense of representationalist construals of predictive processing, see Clark (2013; 2015), Gładziejewski (2016), Williams (2018), and Ryder (2004).

[9] Dan Ryder (2004) stands out as also recognizing the importance of prediction to content attributions, so I take a moment to explain his "SINBAD" theory and contrast it with the forward-looking theory I will sketch below. To begin, Ryder interprets individual cortical pyramidal neurons as mechanisms that can implement a "predictive trick" of discovering mutually correlated clusters of environmental properties by equilibrating their inputs and outputs. Ryder interprets such equilibration as the function of these neurons, and like other teleosemanticists derives the neuron's content from its function. However, Ryder's view on the source of these functions is backward-looking, for he holds that it is determined by the neuron's causal history. Metaphysically, the gulf between this view and the forward-looking view offered here is large, especially regarding the sources of content's normativity. The forward-looking view holds that it is the agent's own forward-directed epistemic capacities that ground a representation's content, rather than historical contingencies concerning the environmental sources of past equilibration. This distance also reflects a disagreement over the explanatory job of representational attributions; on Ryder's view, their primary role is to explain prior equilibration, whereas the forward-looking view holds it is to uncover the epistemic organization of the underlying system and correspondingly to also predict and explain future behavior (including revisions). This strategy works

predictive model; I focus on one of the older and better-understood variants, a model of hippocampal function in associative learning offered by Gluck and Myers (hereafter, the "GM" model)—keeping in mind that predictive learning is likely a governing principle of many other subsystems and that the basic story can be generalized to other kinds of predictive architectures.

The GM model construes the hippocampal system as a type of multi-layer neural network architecture called a predictive autoencoder. I focus on this model for several reasons, all relating to simplicity and clarity of exposition. First, the network's **organization and behavior** is clearer than that of the more abstract, mathematical alternatives, which lack a firm mechanistic interpretation (e.g., Friston 2010). Secondly, what counts as a **vehicle** is clearer in the GM model; our discussion can draw on work by Gärdenfors (2004) and Shea (2007b) specifying the individuation conditions for vehicles in such networks (by contrast, what counts as a vehicle in the more hierarchical or mathematically abstract models remains an open question). Thirdly, what might count as a **content** is clearer in the GM model. Clark, for example, notes that the more Bayesian models appear to encode not traditional concept extensions like *apple* but rather probability density distributions, as well as other features like "uncertainty, noise, and ambiguity" (2013: 8). Clark acknowledges that such distributions (which are perhaps most at home in describing basic perceptual contents) may be quite different from the simpler, sparser, and more univocal set of contents that are ascribed and experienced at the conceptual level—but it remains unclear how the latter is to emerge from the former (though see Wiese 2017). With the GM model, by contrast, we can take advantage of a mature literature—much also due to Clark (1989)—relating such a network's activity to the kinds of contents ascribed in comparative psychology and cognitive ethology. Fourthly, I focus on a predictive-learning system that operates at roughly the whole-organism level that takes high-level inputs from all sensory modalities, rather than various specialized subsystems which might also operate according to predictive principles (such as motor planning). This is because behavior in comparative psychology and ethology is usually explained using whole-organism content ascriptions, as reviewed in the previous section. Finally, the **neuroanatomical realizers** of the GM model are well-understood and empirically-supported, the linkage to the medial temporal lobes being grounded in a great deal of anatomical and cognitive neuroscience, lesion studies, and computational modeling. The hope is that focusing on this older but simpler model will allow us to clarify the metaphysics

because it highlights the ways that the agent itself treats misrepresentation as significant. The worry is that unless we incorporate these revisions into the explanatory purview of the theory, prior equilibration can be explained in terms of past indication relations alone (i.e., a fixed cluster of causal relays—see also Dretske 1986), raising the issues of indeterminacy or restricting us to modest, appearance-like contents.

of forward-looking content without losing the thread in speculative and imprecise interpretations of the newer models.

A predictive autoencoder is a multi-layer neural network for which input and output layers are connected to the same stimulus input stream (Fig. 3). Briefly, the GM model construes the hippocampal region as learning to predict the future state of late-stage, highly-processed, multi-modal stimulus input streaming in from the entorhinal cortex, where the evidence on which the predictions are based is the previous state of that same input stream. Output errors are calculated by computing the distance between the predicted and actual stimulus input and backpropagated to earlier layers to train the network to minimize the likelihood of making similar errors in the future. (Readers concerned with generic heyday arguments against attributing representations to or typing vehicles in such networks are directed to Shea 2007b, which ties together earlier works on this topic, including P. S. Churchland & Sejnowski 1990; Rupert 2001; and Tiffany, 1999.) Many more layers can be added to such a network to increase the degree of compression and hierarchical organization in the network's representations, and indeed the current boom in deep learning was inspired in part by work on deep autoencoders (Hinton & Salakhutdinov 2006).

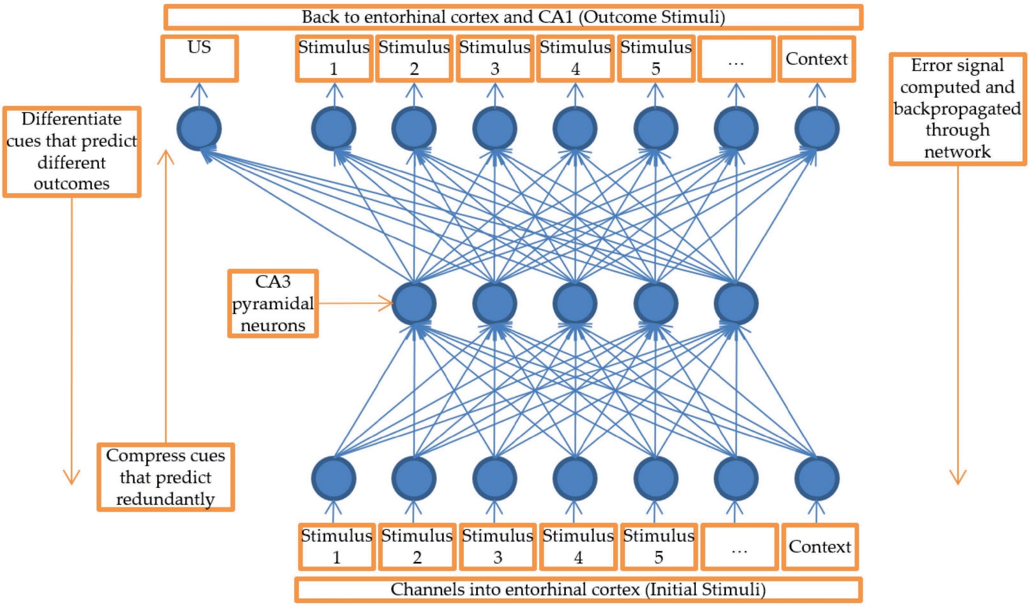
Back to entorhinal cortex and CA1 (Outcome Stimuli)
US
Stimulus 1
Stimulus 2
Stimulus 3
Stimulus 4
Stimulus 5
…
Context
Differentiate cues that predict different outcomes
Error signal computed and backpropagated through network
CA3 pyramidal neurons
Compress cues that predict redundantly
Stimulus 1
Stimulus 2
Stimulus 3
Stimulus 4
Stimulus 5
…
Context
Channels into entorhinal cortex (Initial Stimuli)

**Figure 3.** Gluck and Myers's (1993; 2001) cortico-hippocampal model of configural learning. The US is some unconditioned stimulus (like food), which given its natural connection to unconditioned responses (such as consummatory movements), provides the network with initial behavioral potency. The network's representations can also be made context-sensitive through the use of contextual input/output nodes, understood as unitized configurations of stimuli that reliably identify distinct contexts.

To tie to our earlier discussions, these inputs can be thought of as more basic structures that already possess simpler Tier 1 or 2 contents; now, it is a configuration of features that a representational vehicle's inputs jointly indicate (which I continue to refer to as '*F*', keeping in mind that '*F*' can now also refer to a configuration of proximal indicated stimulus features <*f1*, *f2* . . . *fn*> rather than just a single one). The inputs to the GM system are not coming directly from the world or the sensory organs, but rather from cortical structures that have already been shaped by selective forces to process information about environmental stimuli, whether by evolution, development, or more basic forms of associative learning. The GM network then assembles these Tier 1–2 indicators into configural clusters that can more flexibly predict the future state of that same stimulus input stream. The clusters are revised when they fail to predict the future state of the same indicator-style input stream. Because revision occurs iteratively and discovering the most predictive clusters can take an indefinite amount of time, it is in part the informational structure of the organism's environment (and the organism's active engagement with it) that determines which stimulus patterns will be presented to the network often enough to produce systematic revision. This is why Tier-3 contents cannot be reduced to concatenations of their Tier 1–2 components (i.e., *F*); the interaction between the current structure of a representational cluster and the informational distribution of the environment determines how and when revision will reliably occur, and which revisions will remain stable in the face of future exploration and engagement.

Gluck and Myers describe the revision of this network as striking a balance between two biases, predictive differentiation and redundancy compression. On the one hand, the drive to correctly predict stimulus patterns at the output layer creates a pressure to differentiate stimulus patterns which predict different outcomes, by altering link weights to render their hidden layer activation patterns more distinct.[10] On the other hand, there are fewer hidden layer nodes than input or output nodes, so the system cannot just memorize every statistical regularity in its environment; it must economize on representational resources by treating input stimulus patterns which redundantly predict the same outcomes as being the same, by rendering their hidden-layer activation patterns more similar. The balance between these two pressures (which will be elaborated further in the next section) enables a powerful form of all-purpose inductive learning, which over time will tend to discover an efficient set of stimulus configurations that predict the widest range of stimulus variance in the organism's environment.[11]

[10] The GM model uses error backpropagation learning algorithm to accomplish this; for more discussion of this choice, see Buckner and Garson (2018).

[11] Notably, redundancy compression is the drive which helps to prevent overfitting, a problem facing most machine learning methods. Good generalization performance depends

The learning of such a network can be thought to perform a search for a set of feature-clusters that correspond to attractor basins in multi-dimensional feature space. Attractor basins here mark configurations of input stimuli that indicate clusters of features in the system's environment that tend to reliably co-occur with one another, the joint tracking of which minimizes net prediction error; call the environmental property corresponding to the attractor basin closest to a cluster's current location in state space '*F*+'. In most cases relevant to simple animals, these minima indicate their environments' natural kinds, because these will be the stable sources of feature clustering that could be reliably predicted (Boyd 1991). Once a representation's activation vector brings it sufficiently close to one of these attractor basins, it becomes overwhelmingly likely that if that representation is further revised (in the sense that learning alters the set of proximal features *F* that token it), it will be revised to better indicate *F*+, which will in turn reduce prediction errors and correspondingly future revisions. By contrast, a revision that took the representation further away from its nearest basin would tend to increase prediction error, leading to increased revisions and increased instability.

The upshot is that we cannot rely solely on a configuration's current or past informational relations to identify its Tier 3 contents, since these relations will always be in flux as the contents of the cluster are continuously revised. This provides an overarching reason why scientists should favor forward-looking rather than backward-looking attributions for representations which are subject to revision by prediction-error-reduction mechanisms: to hit a moving target, we should aim where those relations are headed, rather than where they currently are or have previously been. Informally, a representation's forward-looking content (*F*+) in some environment is thus what it indicates at the limit of its likeliest revision trajectory, given that environment's informational structure.[12] More precisely:

upon networks seeking out deeper, more inductively-potent regularities, and redundancy compression's drive to economize on representational resources enforces this drive at the level of the mechanism.

[12] It may help to more precisely define the notion of a revision trajectory. Theories of content can type representations in at least three different ways—by vehicle, mode of presentation, and content—and how a revision trajectory is formally defined will vary depending on how these distinctions are cashed out in a particular model. On the framework suggested here for the GM network, vehicles are clusters of hidden layer activation patterns (Gärdenfors 2004; Shea 2007b), modes of presentation are the configurations of proximal stimuli those clusters currently indicate (if one bristles at the Fregean baggage of "mode of presentation" here, one could substitute "intension", or "stereotype" instead—Putnam 1975), and content (on the here-proffered, Tier 3 forward-looking theory) is the property indicated by the nearest attractor basin that the vehicle is overwhelming likely to be revised to better indicate, if it is revised at all. A revision, in this sense, occurs when a change is made to a hidden-layer activation pattern cluster as a result of a detected prediction error (e.g., by increasing or decreasing the influence of one of its feature components),

> **Forward-looking content:** The (Tier 3) content of a representation *N* currently indicating feature configuration *F* in environment *E* is whatever property *F*+ that (properly functioning) error-correction learning makes systematically likely it will be revised to better indicate in *E*, if it is revised at all.[13]

This forward-looking view jointly satisfies the causal and normative aspects of content attributions. Notably, the detection of misrepresentation, in the form of prediction failure, now does something crucial in the cognitive architecture: it causally shapes how these configurations are revised by the cognitive system over time. The system actively monitors the predictive success of its representational scheme and responds by altering that scheme when it fails to correctly predict important outcomes. Moreover, the trajectory of these revisions is determined by the way the system treats these configurations as representational prediction tools, because they are revised only when they misrepresent in this way. This is true because, by hypothesis, if the organism's representations currently indicated their (forward-looking) contents, their referents and associated features would be present in the stimulus input stream, and there would be no prediction error. Moreover, I have not abstracted away unduly from the actual causal structure of the system in ascribing these more ambitious contents, because the representational attributions which outstrip the agent's current discriminative abilities are grounded in forms of information processing that these creatures can actually perform. A fully naturalizable, empirically-supported story can be told about how evidence of misrepresentation is detected through prediction

but the vehicle remains locked into the same revision trajectory (i.e., its likeliest course through feature space after predictable future revisions terminates in the same attractor basin). Sameness of vehicle across revision is determined by closest activation vector cluster in feature space between time *t* (before revision) and *t+1* (after revision); typically each cluster will have only one nearby counterpart after revision. A revision trajectory for some representation is thus defined as a unified series of predictable, systematic revisions, across which the mixture of proximal features indicated by a hidden-layer activation pattern cluster is continually tweaked to better indicate the same distal property indicated by a stable attractor in feature space.

[13] A reviewer worries about the final clause here, specifically that given the "cobbled-together, kludge-y nature" of cognition, that perhaps nearby backward-looking views that focus on the best explanation for prior stabilization (such as Ryder 2004 or Shea 2018) may have an advantage here. I dig in my heels on this point; I do not think it appropriate to ascribe contents that outstrip even an agent's future discriminative abilities in cognitive science (thus agreeing with arguments offered by Neander 2006 to this effect), for this would prevent us from offering the proposed empirical resolution to the ascriptive disagreements in comparative psychology highlighted in Sections 3 and 5 here. If there are agents that are constitutionally incapable of improving the reliability of their representational strategies even when presented with robust evidence of misrepresentation, then we should not ascribe more ambitious contents than their best-case discriminations can support. The benefit of digging in one's heels here is that the value of Tier 3 content ascriptions can still be justified in terms of benefits offered to the theorist's ability to predict the behavior of the representation's owner; the scope of the prediction targets just needs to be expanded to include responses to evidence of misrepresentation and future behaviors as well.

error, and how representations are revised to better indicate their forward-looking contents as a result.

Two quick and obvious objections are worth confronting at the outset, because they are based on tempting misunderstandings of these claims: the charges of circularity and backwards causation. First, readers may worry here that taking a representation's forward-looking content to be determined by the likeliest limit of prediction-error learning, together with the assertion that prediction error is a form of misrepresentation wherein a representation fails to successfully indicate its forward-looking content, leads to a vicious circularity or regress. This worry would be sound if the prediction-error signals were calculated individually for each representation, that is, if the system were learning by supervised methods wherein an external oracle specified for each trial whether the correct forward-looking contents of any tokened representations were actually present. Lacking an independent account of how the supervisory oracle knew the identity of the forward-looking contents for each representation, I should indeed be stuck in vicious circularity or regress. Secondly and relatedly, the proposed theory would also seem to require a suspicious form of backwards causation: contents whose causal relations are only fully realized after revisions that may occur in the distant future would seem to be involved in computing error signals in the here-and-now. Either of these objections, should they strike home, would admittedly be fatal to the proposed view.

But crucially, the objections of the previous paragraph are based on a misunderstanding of how such systems detect misrepresentation. Prediction errors are assessed not by comparing present predictions to the final outcome of some series of actual future revisions in some specific future timeline; rather, misrepresentation is assessed at each step using the agent's current Tier 1–2 perceptual inputs and current predictions. In other words, after each individual prediction, error is detected by the system not through clairvoyance, but rather by checking the presence or absence of the Tier 1–2 contents predictively associated with that representation by previous learning. For example, it is not as though the system tokens DUCK in the presence of a plastic decoy, and an oracle, which has supernormal access to the future, tells the system, "Sorry, according to my calculations that representation was headed to *duck*, and there's no *duck* here; try again!" It is rather that the system tokens DUCK, and as result of prior learning, predicts that incoming sensory information will also contain (Tier 1–2 features associated with its mode of presentation or intension) *quacking*, *feathers*, *floats on water*, and so on, which have already been associated with DUCK through prior learning.[14] When the system fails to observe some of these expected stimuli, it

[14] It may be worth noting that the representations on this view thus may become more like structural representations than indicator representations over time; though I will not explore this

will realize that it has gotten something wrong, and must make a revision (perhaps it increases the weight of more valid features like *quacking*, relative to the other features which are present in decoys). Crucially, this error calculation does not require or provide information about where the representation is headed in the future, or how exactly it should be revised. The system must discover the stable solutions itself through trial-and-error, and only revisions that reduce prediction error will tend to be retained over time. This explains how Tier-3 representations are bootstrapped from simpler Tier 1–2 components without circularity or backwards causation.

Upon appreciating that the computation of prediction error is imperfect, a third worry might arise: that reducing content to prediction-error correction makes content normatively arbitrary, because any revision that happens to be made as a result of prediction error becomes ipso facto correct. This worry, however, is also based on a misunderstanding. Granted, any particular revision, just like any particular tokening, can of course be mistaken. While the proposed view does hold that every time there is a prediction failure, the system receives evidence that it made a representational mistake, it does not hold that any revision that happens to be made as a result of this error signal ipso facto renders the representation's use more correct (see also Wilson 1982). Because incorrect revisions will generate more prediction errors, a properly-functioning error-correction system will tend to stabilize on more correct use. It does so again without any oracular knowledge as to optimally correct use or when it has ultimately reached it. Correct use is instead determined by the likeliest revision trajectory achieving a stable reduction of prediction error, and again which solutions are ultimately stable for some representation is determined by ongoing interaction with the informational structure of its environment, rather than the moment-by-moment fluctuations of its learning mechanisms. This is why systems can possess forward-looking contents before learning completes (and even if learning never definitively finishes). Learning must instead only progress to the point where a representational cluster crosses the threshold of influence of a nearby stable attractor basin, by having enough appropriate Tier 1–2 features associated with it to "point it in the right direction".

I have thus dealt with some of the most obvious objections to this kind of forward-looking view, though there remain many details in need of explication. In particular, I could clarify the sense of "indication" that applies at the end of these revision trajectories. Fortunately, the forward-looking gambit works almost however we come down on these issues; I could perhaps set aside Dretske's much-criticized requirement of perfect correlation (i.e., $P(F|N) = 1$—see Slater

point further here since the upshot of recent literature on this topic is that this distinction is one of degree and not of kind—see Morgan (2014), Nirshberg and Shapiro (2021)

1994) and endorse one of the more recent informational criteria of Rupert (1999), Usher (2001), Eliasmith (2005), or Scarantino (2015), which can ascribe plausible contents with less stringent measures even at the limit of a revision trajectory. The differences between these views become less pressing once the range of eligible contents has been winnowed to stable attractors, and learning has honed informational contact to the limit of its ability.

## 5. Explanations and Examples

It may help to illustrate the theory's implications by returning to some of the themes from Section 2 and 3, and by showing how the view would tackle some familiar and novel cases. To review, I have now provided a theory of content for Tier 3 configurations. A sufficiently sophisticated predictive learning system can bootstrap Tier 3 representational content from Tier 1–2 indicators by flexibly reconfiguring those representations in response to evidence of prediction error. Prediction-learning systems continually monitor the fit between their associated predictions and ongoing observations, and revise these structures where there is a mismatch. Moreover, on the theory provided here, these Tier 3 representations have contents that outstrip the sum of their Tier 1–2 components; the contents are forward-looking, because the terminus of a learning trajectory cannot be reduced to a mere concatenation of its Tier 1–2 components, which will be undergoing constant revision. Any stability to be found in this learning process is rather determined in part by the locations of the attractor basins that would minimize prediction error after likely future revisions.

To illustrate the implications of this view, let us compare two different types of cognitive systems, starting with the classic "disjunction problem" example of the frog whose tongue darts out in the presence of flies and beebees. Does the neural structure which controls these tongue-darting movements have the ambitious content *fly*, a disjunctive content like *fly* ∨ *beebees*, or some modest conjunction of proximal stimuli like *small, dark, moving speck*? Suppose that the frogs, as Neander (2006) has argued on the basis of neuroethological research, respond inflexibly to artificial lures like *beebees*, and further that these frogs are not capable of altering their behaviors in response to repeated evidence that the stimuli to which they respond are not in fact flies. On the theory supplied here, the frogs simply lack Tier 3 contents; the neural structure controlling tongue-darting movements is not under the control of a learning system which continually improves its ability to track *flies*. The same, moreover, would apply to magnetosomic bacteria and many other counterexamples from the heyday literature.[15] If these organisms cannot revise their indicator states in response to evidence of error, then they do not count as having Tier 3 contents at all.

Fodor drives this point home in a prescient comment on the difference between humans and frogs:

> The relevant consideration isn't however, just that frogs sometimes go for bee-bees; it's that they are prepared to go on going for bee-bees forever. Sometimes I swat at mere fly-appearances; but usually I only swat if there's a fly. Sometimes Macbeth starts at mere dagger appearances; but most of the time he startles only if there's a dagger. What Macbeth and I have in common—and what distinguishes our case from the frog's—is that though he and I both make mistakes, we are both in a position to recover. By contrast, frogs have no way at all of telling flies from bee-bees. (Fodor 1990: 240).[16]

To repurpose the point, it is a reliable fact about our more sophisticated cognitive architecture that if our representations are really about *fly* and not *fly* ∨ *beebees*, or really about *dagger* and not merely *dagger-like-appearances*, then were we repeatedly exposed to beebees or dagger-imposters and allowed to learn about the consequences of our mistaken reactions to them, we should have some capacity to revise our representations so as to no longer respond to these misleading situations. Agents that systematically persist in committing the same kinds of errors forever should not be counted as possessing the more ambitious contents. In short, the forward-looking mantra is something like: fool me once, shame on you; fool me indefinitely, shame on my contents.

To illustrate the significance of this kind of generalization to cognitive science, let us now consider a class of more cognitively flexible animals that share this capacity to revise: the chimpanzees, monkeys, and corvids that were the subject of interpretive stalemate in the Theory of Mind literature discussed in Section 3. The problem was that prior experiments were unable to determine whether animals ambitiously represent *seeing* (i.e., *F*+) or more modestly just the proximal stimuli that indicate seeing, like *line-of-gaze* (i.e., *F*). The forward-looking strategy suggests a way to overcome this impasse; animals that really represent *seeing* should be able to recruit an open-ended variety of other non-gaze cues to better indicate it, flexibly generalizing predictions from old to new cues in a way that cannot be reduced to a fixed set of more proximal features (for a

[15] See the discussion in Dretske (1986) for a similar treatment of magnetosomic bacteria.

[16] For a similar point, see Bielecka and Miłkowski (2019)—though while they simply require that the agent be able to detect the error, I additionally require that it be disposed to revise the representation to make the error less likely in the future.

summary of the numerous proximal cues <*f1*,*f2* . . . *fn*> that have been invoked by skeptics under the heading *line-of-gaze*, see Fletcher & Carruthers 2012).

## 6. The Causal Relevance of Forward-Looking Content

To see the distinctive power of forward-looking content in causal explanations of behavior, it will help to further explore the two forces of representational revision discussed above: predictive differentiation and redundancy compression. These forces act on cue configurations rather than the sort of elemental cue-behavior links characteristic of behaviorism, and they reshape those cue configurations in ways that are determined by their predictive value. It will help to explore each force in turn, and to illustrate how backward-looking views fail to deliver the correct content ascriptions in cases where predictive differentiation or redundancy compression operate.

The more straightforward of the two revision forces is predictive differentiation. William James highlighted the significance of predictive differentiation in shaping the mind's representations in a famous passage where he describes learning to distinguish claret from Burgundy. James reports that although these two wines initially tasted almost identical to him, with practice he learned to emphasize the very slight differences amongst their taste profiles, because only these differences were useful in predicting their correct labels. He writes that the forces which revised these representations acted upon them as wholes:

> The effect of practice in increasing discrimination must then, in part, be due to the reinforcing effect, upon an original slight difference between the terms, of additional differences between the diverse associates which they severally affect. (James 1890: 511)

He noted that the effect of these forces was to pull the two representations further apart in perceived similarity, so that "small differences affect us as if they were large ones" (1890: 515). Causally, we may suppose that these forces were powered by the reinforcing effect of attempting to predict the correct labels, acting upon cue configurations as wholes. Whenever the wine's label is predicted incorrectly, whichever components of the configuration were attended to in the misclassification (e.g., the acidity of a claret) will tend to have its strength weakened in the next revision of that configuration; and whenever the wine's label is predicted correctly, the components that led to the correct classification (e.g., the tannic quality in a Burgundy) will tend to see their strength correspondingly increased relative to the other components in that configuration. The reshaping leads over time to more reliable classifications; though James might originally have been at chance in distinguishing the two French styles of wine, with enough diligent drinking—in the name of science, of course—he reports eventually succeeding at the task.

Predictive differentiation is probably the most common form of prediction-error-driven revision, and the easiest place to see the importance of a forward-looking theory of content for a wide range of cases. The forward-looking view matches our intuitions about a variety of commonplace cases where predictive differentiation operates, especially concept learning in young children. When we took my three-year-old daughter to the zoo to see a tiger for the first time, like many other children her age, she exclaimed "Look at the big kitty!" We must here say that she lacks a representation of *cat* entirely (perhaps representing instead *cat* ∨ *tiger* ∨ *lion* ∨ *dog on a dark night* ∨ . . .), or that she somehow possesses a representation of that property even when she makes the mistake, despite never having had a better ability to distinguish cats from tigers earlier in her life. Since she possessed a great deal of experience with domestic cats, including features acquired from this experience such as that cats have fur, whiskers, a tail, and so on—features that causally explained why she made the mistake she did in this case—saying that she lacks a representation of *cat* entirely seems inappropriate, and leaves us unable to explain her mistake in the natural manner, by saying that she thought the tiger looked like a cat. And since we reliably predict that she will, like most other children, eventually come to distinguish cats from tigers upon learning of her mistake—by better attending to and appropriately weighing differences which initially seemed slight, as James suggested—we should say that her representation means *cat* even when she makes her mistake and is still in the process of learning. It is even clearer that this is the correct answer when we see that the kind of mistake she makes differs only in degree from those that remain in even adult use (e.g., perhaps upon the first time we see an ocelot, fossa, or civet).[17]

[17] A reviewer wonders whether it might be just as likely that the child's representation gets revised to better indicate *tiger* to the exclusion of *cat*. Depending upon the vehicle's current indicator properties and the environment in which it would be revised, this is certainly possible; what matters is which revision is, as a matter of fact in that environment, overwhelmingly more likely to occur. Other outcomes are also possible; the representation could be poised midway between attractor basins corresponding to *cat* and *tiger*, and the likeliest outcome is that it gets "split" in two by revision, in a form of representational mitosis (perhaps this is so at age 3, but not by age 5, for example). It is also possible that a representation could, as a matter of fact, undergo a series of low-probability revisions that knock it out of the influence of one attractor basin and into another; this would reflect a change of content at some point in the (highly unlikely) trajectory. Rather than considering such questions to be problems for my view, they are gleefully endorsed as avenues for the research program to explore. Indeed, I think if we step away from prior intuitions about teleosemantics for a moment, these are the right questions to be asking, mirroring those about semantic change and deference to future use that have long featured prominently in philosophy of language (see, e.g., Ebbs 2009 for a review and discussion). That such questions arise naturally

James also noted the significance of the other predictive force shaping our representational scheme, redundancy compression. Indeed, he also laid it down as a basic principle of the mind's organization, claiming that "any total impression made on the mind must be unanalyzable, whose elements are never experienced apart" (1890: 502). In other words, we should find it difficult to even discriminate two different stimuli unless they at some point occur in different circumstances or predict different outcomes. Redundancy compression is thus intimately related with predictive differentiation, in that it both serves as the mind's default state which predictive differentiation must overcome to make two cue configurations discernible, and continuously serves as a headwind against the wanton use of representational resources, pushing representations back to an original state of indiscernibility if their distinction does not deliver predictive goods. "If all cold things were wet and all wet things cold," James wrote, "if all hard things pricked our skin and no other things did so; is it likely that we should discriminate between coldness and wetness, and hardness and pungency respectively?" (1890: 502). James provides several actual examples of such co-occurring and initially indiscernible phenomena—the contraction of the diaphragm and the expansion of the lungs, or the convergence of the eyes and focusing on an object—noting that it is only with the aid of theory that we come to represent these things distinctly at all, by establishing distinct causes and effects that allow us to imagine them occurring apart.

Returning to the GM model of hippocampal learning, we can provide a mechanistic explanation for redundancy compression, as it emerges from the informational bottleneck in the network's hidden layer imposes on predictive differentiation. The key insight is that when there are more configurations which must be distinguished than there are nodes to represent them one-for-one in the hidden layer, representations must compete with one another for common resources. As one representation is revised in response to predictive differentiation, it will "jostle" the weights of distinct but related cue configurations which share some of their elements. This will in turn cause an increase in errors involving tasks drawing upon the other representations, which may suffer from the revisions as innocent bystanders. The only way for a network to minimize net representational error across its whole suite of representations is to find opportunities to economize, by compressing distinct configurations that predict redundant outcomes into the same representation. The fewer distinct representations the network has to maintain, the more resources are available to other representations, and consequently the fewer bystanders become casualties in subsequent predictive differentiation. As the co-predicting representations

within the forward-looking framework proposed is a benefit of the view, though a full exploration of the different kinds of cases and their appropriate verdicts must be left to future work.

gradually come to coincide, the creature so revising its representations will also enjoy the advantages of generalizing the information it has learned about one representation to the others, as a free by-product of predictive differentiation being forced to operate under the need for efficiency. Notably, this will tend to also render representations more inter-subjective, reducing differences derived from idiosyncrasies in learning histories from agent to agent, so long as they live in a shared environment.

Many cases involving redundancy compression call for more rapid revisions, however, suggesting that it may take two different forms: the slow and passive return to stable indistinguishability just described, and a more active form that seeks out opportunities for economy and missed generalization, which result in rapid revisions once discovered. Rapid revisions are more appropriate for many of the so-called "twin" cases discussed in the heyday literature, wherein one content is represented by two distinct concepts; to return to another stock heyday example (Fodor 1994), Oedipus did not require a long period of gradual prediction failure to realize that Jocasta was actually his mother, instead performing some quick revisions which resulted in some important new generalizations with narratively significant consequences. However, adding a degree of surprise to the model which modulates the learning rate of the network can show how both more passive and active forms of redundancy compression could be implemented in the same mechanism.

A rich body of empirical research has demonstrated that the mammalian brain possess such mechanisms to regulate its learning rate to suit circumstance; research has especially focused on the role of acetylcholine in modulating the plasticity of the hippocampus and cortex in learning (Hasselmo 2006). Acetylcholine is a neurotransmitter which is released by the nucleus basilis into the hippocampal and cortical tissues in response to surprisal. Among its other functions, it increases the plasticity of these tissues for a short period of time. This can produce a period of rapid revision that subsides when the source of the prediction failure is corrected and surprisal diminishes. Indeed, the entire plot of Sophocles' play revolves around Oedipus confronting a series of improbable and unpredictable calamities: that the man he killed on the road years ago was the former Theban king Laius, that as a baby he was delivered to Corinth by a shepherd instead of born to Corinth's king Polybus, and that his wife had ordered the shepherd to kill Oedipus as a baby so as to avoid the fulfillment of a prophecy that the child would murder Laius. Anyone in his position at that moment would have been so pumped full of acetylcholine that they could quickly compress their representations for JOCASTA and MAMA (learning the truth about their tragic marriage would be the complimentary "free" generalization). Instead of two different forms of redundancy compression, we may thus really have only a single continuum, which can be modulated by altering the rate at which the learning takes place in response to the degree of surprise that the prediction failures generate.

In considering the empirical relevance of forward-looking contents, this more rapid form of redundancy compression is of particular interest, because it could enable the leaps of "insight" that animal cognition researchers often seek to demonstrate in their experimental designs. Different kinds of agents may differ in their ability to recognize when two perceptually diverse situations *A* and *B* predict some of the same outcomes; particularly intelligent and attentive animals may excel at this ability, using it to quickly compress two representations into a single vehicle in the course of one experiment. This will in turn cause an animal to automatically generalize all other behaviors and predictions previously associated with *A* to *B* and vice versa. As a result, the subject will appear to demonstrate "insight" in its behaviors; if the animal previously learned that cue *A* predicted outcome *X*, and later learns that *A* and *B* both predict some other shared outcomes $Y_1$, $Y_2$ . . . $Y_n$ it can now "infer" that cue *B* also predicts *X*, despite the fact that *B* and *X* never actually co-occurred anywhere in its learning history.

An experiment by Bugnyar, Reber, and Buckner (2016) purports to provide just such a demonstration of representational flexibility in ravens (Figure 3), in a way which engages with the interpretive debate over Theory of Mind which was reviewed in Section 3. To provide some background, ravens spontaneously cache food items as part of their normal foraging behavior, and like chimpanzees, previous experiments have shown that they behave differently when they are being watched by a competitor (who might pilfer their caches). Specifically, they cache more quickly, are less likely to return to previous caches, and have even been observed to make "false caches"—poking their beaks in the ground while retaining the food in their throat pouch (Bugnyar 2013). However, in all prior designs the ravens could see the competitor's gaze cues at test, and so this body of research had until recently faced the same interpretive stalemate as the research on chimpanzees (Heyes 2015).

The recent experiment from Bugnyar's lab finally overcame this problem by showing that ravens could recruit a new, non-gaze cue for *seeing*, and use it to draw novel inferences (Fig. 4).[18] Specifically, this experiment provided evidence that ravens trained to use peepholes to pilfer another's caches can later infer that when they cache in the presence of an open peephole—even if they had no prior experience caching in the presence of peepholes—that unseen competitors might be able to watch them through the peephole, too. As a result, they later guard their own caches against observation in the presence of the peephole, even if they

[18] This is not to say that all the skeptics have been convinced; see Lurz (2017) and Kuznar, Pavlic, Glorioso, and Povinelli (2020).

cannot see any competitors (or their gaze cues) at the time. A plausible explanation of this result is that they are highly attentive to their own pilfering opportunities, and were (pleasantly) surprised by the discovery that the peephole afforded them the ability to pilfer the experimenter's caches. This in turn allowed them to learn that two perceptually dissimilar situations—<A sees B out in the open> and <A sees B through a peephole>—predict some of the same significant outcomes, and so they compressed their representations of these two situations into an "abstract equivalence class" that economized on their representational resources (Whiten 1996). As a side effect of this revision, they were able to generalize an independent prediction associated with a food item's being visible with the peephole cue. In other words, the compression of the representations of the two states of affairs would automatically grant them the "insight" that others be able to pilfer their caches by watching them through the peephole, too.

This kind of revision constitutes a form of representational control that significantly outstrips the less-flexible forms of learning explored by other heyday informational views. The attribution of the content *seeing* over *line-of-gaze* crucially allowed the researchers to make this prediction (and to explain the successful outcome), because there are no gaze cues actually present in the test condition of the experiment. This demonstrates that not only was *seeing* (i.e., the *F*+ candidate) the right content to attribute to the ravens all along—even before the recruitment of the peephole cue, when the ravens' representations only synchronically indicated *line-of-gaze* (i.e., *F*)—but also that only the forward-looking interpretation of this content attribution could provide the resources to devise this experiment and explain its results, thereby breaking this longstanding empirical stalemate.[19]

Though I have here focused on the base case of Tier 3 representations—explaining how they are configured and derived from Tier 1–2 contents—a similar idea could be extended to more elaborate still "Tier 4" representations, which might be scaffolded by more metacognitively sophisticated and socially-regulated forms of representation. For example, Shea, Boldt, Bang, Yeung, Heyes, and Frith (2014) have recently proposed a two-systems account of metacognition that provides a distinctive and important role for interpersonal communication in achieving more social and explicit forms of cognitive control over representational revisions. Organisms that could explicitly consider their representational successes and communicate about them with one another would achieve forms of representational and behavioral flexibility unavailable by other means. They could accumulate and transmit a shared conceptual and

[19] This was only one experiment, but a structurally similar experiment with a similar outcome has been performed in chimpanzees, with the learned cue being the color of boxes with unseen objects inside—see Karg, Schmelz, Call, & Tomasello (2016).

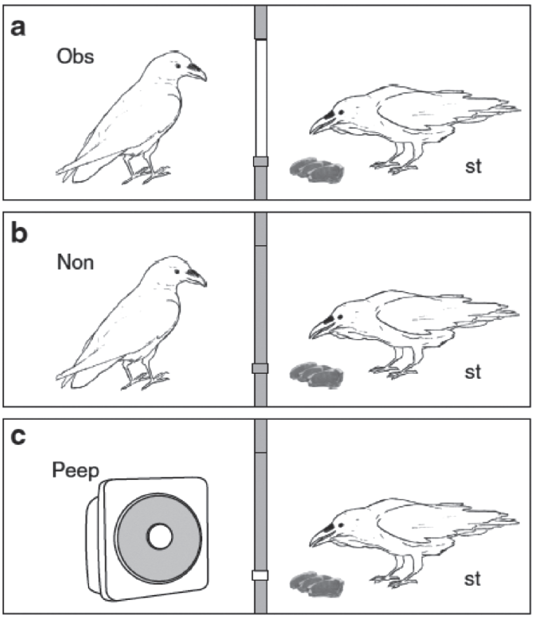

**Figure 4.** Sketch of the experimental setup of Bugnyar et al. (2016). (a) Observed (Obs) condition: The cover of the window is open (white bar) and the focal subject (storer, st) caches food in the visual presence of a conspecific (observer). (b) Non-observed (Non) condition: The cover of the window is closed (grey bar) and the focal subject caches food in visual isolation of a conspecific (non-observer). Both observers and non-observers make sounds in the experimental chamber, which are audible to the storer. (c) Peephole (Peep) condition: The cover of the window is closed (grey bar) but one of the two peepholes (small white square) is open; the focal subject caches food in the absence of any behavioral cues, whereas the presence of conspecifics is simulated via playback of sounds recorded from non-observed trials (symbolized by loudspeaker). In this experiment, the storer raven was first trained to use peepholes to pilfer caches it observes being made by an experimenter. The Obs and Non conditions were used to see how the raven behaved when it knew it was and was not being observed, respectively. Caching behavior in the Peep condition—the first time the raven had ever cached in the presence of the peephole—was then found to match that in the Obs condition and to differ from that in the Non condition, despite the fact that no gaze cues were present.

cultural knowledge to one another over generations, and reflectively evaluate their epistemic successes or failures together. This could in turn explain how Tier 3 agents bootstrap a further kind of Tier 4 content from humbler Tier 3 beginnings—which might connect us with more familiar linguistic, theory-mediated accounts of reference to natural kinds, such as those of Quine, Putnam, and Boyd (Boyd 1999; Putnam 1975; Quine 1969; and see also Wilson 1982). This strategy might not only fill in a missing link between deflationary informational theories of content and inflationary approaches to distinctively human intentionality, but also provide a fully naturalized story about how each stage bootstraps the next without either over- or under-intellectualizing each Tier of capacities (Beisecker 1999; Huebner 2011).

## 7. Conclusion

By updating one of the most popular heyday accounts, the forward-looking view here offers a theory of content with distinctive strengths. It grounds at least one kind of forward-looking content (Tier 3) in the limit of revision trajectories determined not only by a representations' past or current causal relations, but also in the way suitably flexible learning systems are disposed to revise those representations in that environment. As a result, it can often assign contents to whole-agent-level representations which are more ambitious than views which focus on past or synchronic informational relations alone. Furthermore, the ability of such systems to so revise is determined by well-understood capacities to detect evidence of misrepresentation through prediction failure, by bootstrapping upon predictions from simpler (Tier 1–2) contents. And finally, the theory can harmonize these tensions while providing useful guidance for the design of cutting-edge behavioral experiments, addressing pressing empirical debates in psychology and ethology. Specifically, when indeterminacy worries arise for these ambitious contents, the agent's own behavior can be used to resolve the ascriptive disagreements, by presenting the animal with evidence of misrepresentation and seeing whether it revises its representations to better track the more ambitious contents in response. Thus, contrary to the pessimists, the prospects for content naturalism's future are good.

## Acknowledgments

This paper has been in development for some time, and space does not permit me to thank individually everyone who offered comments and guidance during its journey. I do want to especially thank the Southern Society for Philosophy and Psychology which awarded this paper with the Griffith Memorial Award, the Alexander von Humboldt Foundation and the National Science Foundation (Grant #2020585) for financial support, a conscientious reviewer for this journal for helpful comments, and especially Fred Dretske, for being so generous with careful, voluminous, and kind correspondence throughout his final years, even with a graduate student he had never met in person and who stubbornly insisted that his views on content were closer to right in 1986 than in 1988.

## References

Allen, Colin (1999). Animal Concepts Revisited: The Use of Self-Monitoring as an Empirical Approach. *Erkenntnis*, *51*(1), 537–44. https://doi.org/10.1023/A:1005545425672

Beisecker, David (1999). The Importance of Being Erroneous: Prospects for Animal Intentionality. *Philosophical Topics*, *27*(1), 281–308.

Bickhard, Mark (1993). Representational Content in Humans and Machines. *Journal of Experimental & Theoretical Artificial Intelligence*, *5*(4), 285–333.

Bielecka, Krystyna and Marcin Miłkowski (2019). Error Detection and Representational Mechanisms. In J. Smortchakova, K. Dołrega, and T. Schlicht (Eds.), *What Are Mental Representations?* (1–31). Oxford University Press.

Botvinick, Matthew, Sam Ritter, Jane X. Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis (2019). Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, *23*(5), 408–22.

Boyd, Richard (1991). Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds. *Philosophical Studies*, *61*(1), 127–48.

Boyd, Richard (1999). Kinds, Complexity and Multiple Realization. *Philosophical Studies*, *95*(1), 67–98.

Brigandt, Ingo (2005). The Instinct Concept of the Early Konrad Lorenz. *Journal of the History of Biology*, *38*(3), 571–608.

Buckner, Cameron (2011). Two Approaches to the Distinction between Cognition and "Mere Association". *International Journal of Comparative Psychology*, 24(4), 314–48.

Buckner, Cameron (2014). The Semantic Problem(s) with Research on Animal Mind-Reading. *Mind & Language*, *29*(5), 566–89.

Buckner, Cameron (2015). A Property Cluster Theory of Cognition. *Philosophical Psychology*, *28*(3), 307–36.

Buckner, Cameron and James Garson (2018). Connectionism and Post-Connectionist Models. In M. Sprevak and M. Columbo (Eds.), *The Routledge Handbook of the Computational Mind* (76–91). Routledge.

Bugnyar, Thomas (2013). Social Cognition in Ravens. *Comparative Cognition & Behavior Reviews*, *8*, 1–12.

Bugnyar, Thomas, Stephan A. Reber, and Cameron Buckner (2016). Ravens Attribute Visual Access to Unseen Competitors. *Nature Communications*, *7*, 10506.

Call, Josep and Michael Tomasello (2008). Does the Chimpanzee Have a Theory of Mind? 30 Years Later. *Trends in Cognitive Sciences*, *12*(5), 187–92. https://doi.org/10.1016/j.tics.2008.02.010

Chemero, Anthony (2011). *Radical Embodied Cognitive Science*. MIT Press.

Churchland, Patricia Smith and Terrence J. Sejnowski (1990). Neural Representation and Neural Computation. *Philosophical Perspectives*, *4*, 343–82.

Churchland, Paul M. (1981). Eliminative Materialism and Propositional Attitudes. *The Journal of Philosophy*, *78*(2), 67–90.

Clark, Andy (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing* (Vol. 6). MIT Press.

Clark, Andy. (2013). Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science. *Behavioral and Brain Sciences*, *36*(3), 181–204.

Clark, Andy (2015). *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*. Oxford University Press.

Cummins, Robert (1990). Meaning and Mental Representation. *Mind*, *99*(396), 637–42.

Cummins, Robert, Hilary Putnam, and Ned Block (1996). *Representations, Targets, and Attitudes*. MIT Press.

Dennett, Daniel C. (1983). Intentional Systems in Cognitive Ethology: The "Panglossian Paradigm" Defended. *Behavioral and Brain Sciences*, *6*(3), 343–90.

Dickie, Imogen. (2015). *Fixing reference*. Oxford University Press.

Dretske, Fred (1981). Knowledge and the Flow of Information. MIT Press.

Dretske, Fred (1986). Misrepresentation. In R. Bogan (Ed.), *Belief: Form, Content and Function* (17–36). Clarendon Press.

Dretske, Fred (1988). *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

Ebbs, Gary (2009). *Truth and Words*. Oxford University Press.

Eliasmith, Chris (2005). Neurosemantics and Categories. In H. Cohen and C. Lefebvre (Eds.), *Handbook of Categorization in Cognitive Science* (1035–54). Elsevier.

Ericsson, K. Anders and Jacqui Smith (1991). *Toward a General Theory of Expertise: Prospects and Limits*. Cambridge University Press.

Favorov, Oleg V. and Dan Ryder (2004). SINBAD: A Neocortical Mechanism for Discovering Environmental Variables and Regularities Hidden in Sensory Input. *Biological Cybernetics*, *90*(3), 191–202.

Fletcher, Logan and Peter Carruthers (2012). Behavior-Reading versus Mentalizing in Animals. In J. Metcalfe and H. Terrace (Eds), *Agency and Joint Attention* (p–p). Oxford University Press.

Fodor, Jerry A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.

Fodor, Jerry A. (1990). *A Theory of Content and Other Essays*. MIT Press.

Fodor, Jerry A. (1994). *The Elm and the Expert: Mentalese and Its Semantics*. MIT Press.

Friston, Karl (2010). The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, *11*(2), 127–38. https://doi.org/10.1038/nrn2787

Gärdenfors, Peter (2004). *Conceptual Spaces: The Geometry of Thought*. MIT Press.

Gershman, Samuel J. (2018). The Successor Representation: Its Computational Logic and Neural Substrates. *Journal of Neuroscience*, *38*(33), 7193–200.

Gładziejewski, Paweł (2016). Predictive Coding and Representationalism. *Synthese*, *193*(2), 559–82.

Gluck, Mark A. and Catherine E. Myers (1993). Hippocampal Mediation of Stimulus Representation: A Computational Theory. *Hippocampus*, *3*(4), 491–516.

Gluck, Mark A. and Catherine E. Myers (2001). *Gateway to Memory: An Introduction to Neural Network Modeling of the Hippocampus*. MIT Press.

Godfrey-Smith, Peter (1992). Indication and Adaptation. *Synthese*, *92*(2), 283–312.

Godfrey-Smith, Peter (2014). On Folk Psychology and Mental Representation. In H. Clapin, P. Staines, and P. Slezak (Eds.), *Representation in Mind: New Approaches to Mental Representation* (147–62). Elsevier.

Grush, Rick (2004). The Emulation Theory of Representation: Motor Control, Imagery, and Perception. *Behavioral and Brain Sciences*, *27*(3), 377–96. https://doi.org/10.1017/S0140525X04000093

Hare, Brian, Josep Call, B. Agnetta, and Michael Tomasello (2000). Chimpanzees Know What Conspecifics Do and Do Not See. *Animal Behaviour*, *59*(4), 771–85.

Hasselmo, Michael (2006). The Role of Acetylcholine in Learning and Memory. *Current Opinion in Neurobiology*, *16*(6), 710–15.

Heyes, Cecilia (2015). Animal Mindreading: What's the Problem? *Psychonomic Bulletin & Review*, *22*(2), 313–27.

Hinton, Geoffrey E. and Ruslan R. Salakhutdinov (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*(5786), 504–7.

Hohwy, Jakob (2013). *The Predictive Mind*. Oxford University Press.

Huebner, Bryce (2011). Minimal Minds. In T. Beauchamp and L. G. Frey (Eds.), *Oxford Handbook of Animal Ethics* (441–68). Oxford University Press.

Hutto, Daniel D. and Erik Myin (2012). *Radicalizing Enactivism: Basic Minds without Content*. MIT Press.

James, William (1890). *Principles of Psychology*. Henry Holt & Co.

Karg, Katja, Martin Schmelz, Josep Call, and Michael Tomasello (2016). Differing Views: Can Chimpanzees Do Level 2 Perspective-Taking? *Animal Cognition*, *19*(3), 555–64.

Kulkarni, Tejas D., Ardavan Saeedi, Simanta Gautam, and Samuel J. Gershman (2016). Deep Successor Reinforcement Learning. *ArXiv Preprint*. *ArXiv:1606.02396*

Kuznar, Shannon, Mateja Pavlic, Gabrielle Glorioso, and Daniel Povinelli (2020). Deconstructing the Raven's Theory of Mind: An Analysis of Bugnyar et al. (2016). *Animal Behavior and Cognition*, *7*(4), 653–57.

Luc, Pauline, Natalia Neverova, Camille Couprie, Jakob Verbeek, and Yann LeCun (2017). Predicting Deeper into the Future of Semantic Segmentation. In *Proceedings of the IEEE International Conference on Computer Vision* (648–57).

Lurz, Robert (2011). *Mindreading Animals: The Debate over What Animals Know about Other Minds*. MIT Press.

Lurz, Robert (2017). Animal Mindreading: The Problem and How It Can Be Solved. In J. Beck and K. Andrews (Eds.), *The Routledge Handbook of Animal Minds* (229–37). Taylor & Francis.

Mendelovici, Angela A. (2018). *The Phenomenal Basis of Intentionality*. Oxford University Press.

Mendelovici, Angela and David Bourget (2014). Naturalizing Intentionality: Tracking Theories versus Phenomenal Intentionality Theories. *Philosophy Compass*, *9*(5), 325–37.

Miłkowski, Marcin (2015). Satisfaction Conditions in Anticipatory Mechanisms. *Biology & Philosophy*, *30*(5), 709–28.

Millikan, Ruth Garrett (1984). *Language, Thought, and Other Biological Categories*. MIT Press.

Millikan, Ruth Garrett (1989). Biosemantics. *The Journal of Philosophy*, *86*(6), 281–97.

Morgan, Alex (2014). Representations Gone Mental. *Synthese*, *191*(2), 213–44.

Nanay, Bence (2014). Teleosemantics without Etiology. *Philosophy of Science*, *81*(5), 798–810.

Neander, Karen (2006). Content for Cognitive Science. In G. Macdonald and D. Papineau (Eds.), *Teleosemantics* (167–94). Oxford University Press.

Neander, Karen (2017). *A Mark of the Mental: In Defense of Informational Teleosemantics*. MIT Press.

Nirshberg, Gregory and Lawrence Shapiro (2021). Structural and Indicator Representations: A Difference in Degree, Not Kind. *Synthese*, *198*, 7647–64.

Papineau, David (1988). *Reality and Representation*. Blackwell.

Penn, Derek C. and Daniel J. Povinelli (2007). On the Lack of Evidence that Non-Human Animals Possess Anything Remotely Resembling a 'Theory of Mind. *Philosophical*

*Transactions of the Royal Society of London – Series B: Biological Sciences, 362*(1480), 731–44.

Prinz, Jesse J. (2000). The Duality of Content. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 100*(1), 1–34.

Putnam, Hilary (1975). The Meaning of 'Meaning'. In K. Gunderson (Ed.), *Language, Mind, and Knowledge*. (139–91). University of Minnesota Press.

Quine, Willard V. O. (1969). Natural Kinds. In N. Rescher (Ed.), *Essays in Honor of Carl G. Hempel* (5–23). Riedel.

Ramsey, William, Stephen Stich, and Joseph Garon (1990). Connectionism, Eliminativism and The Future of Folk Psychology. *Philosophical Perspectives, 4*, 499–533.

Ramsey, William M. (2007). *Representation Reconsidered*. Cambridge University Press.

Ristau, Carolyn A. (1991). Aspects of the Cognitive Ethology of an Injury-Feigning Bird, the Piping Plover. In C. A. Ristau (Ed.), *Cognitive Ethology: The Minds of Other Animals; Essays in Honor of Donald R. Griffin* (91–126). Lawrence Erlbaum.

Rowlands, Mark (1997). Teleological Semantics. *Mind, 106*(422), 279–303.

Rupert, Robert D. (1999). The Best Test Theory of Extension: First Principle(s). *Mind & Language, 14*(3), 321–55.

Rupert, Robert D. (2001). Coining Terms in the Language of Thought: Innateness, Emergence, and the Lot of Cummins's Argument against the Causal Theory of Mental Content. *The Journal of Philosophy, 98*(10), 499–530.

Rupert, Robert D. (2011). Embodiment, Consciousness, and the Massively Representational Mind. *Philosophical Topics, 39*(1), 99–120.

Rupert, Robert D. (2018). Representation and Mental Representation. *Philosophical Explorations, 21*(2), 204–25.

Ryder, Dan (2004). SINBAD Neurosemantics: A Theory of Mental Representation. *Mind & Language, 19*(2), 211–40.

Scarantino, Andrea (2015). Information as a Probabilistic Difference Maker. *Australasian Journal of Philosophy, 93*(3), 419–43.

Seth, Anil K. (2014). The Cybernetic Bayesian Brain. *Open MIND*. MIND Group. Retrieved from https://open-mind.net/papers/the-cybernetic-bayesian-brain

Shea, Nicholas (2007a). Consumers Need Information: Supplementing Teleosemantics with an Input Condition. *Philosophy and Phenomenological Research, 75*(2), 404–35.

Shea, Nicholas (2007b). Content and Its Vehicles in Connectionist Systems. *Mind Language, 22*(3), 246–69. https://doi.org/10.1111/j.1468-0017.2007.00308.x

Shea, Nicholas (2012). Reward Prediction Error Signals Are Meta-Representational. *Noûs, 48*(2), 314–41. https://doi.org/10.1111/j.1468-0068.2012.00863.x

Shea, Nicholas (2013). Naturalising Representational Content. *Philosophy Compass, 8*(5), 496–509.

Shea, Nicholas (2018). *Representation in Cognitive Science*. Oxford University Press.

Shea, Nicholas, Annika Boldt, Dan Bang, Nick Yeung, Cecilia Heyes, and Chris D. Frith (2014). Supra-Personal Cognitive Control and Metacognition. *Trends in Cognitive Sciences, 18*(4), 186–93.

Sims, Andrew (2017). The Problems with Prediction: The Dark Room Problem and the Scope Dispute. *Open MIND*. MIND Group. Retrieved from https://predictive-mind.net/papers/the-problems-with-prediction

Slater, Carol (1994). Discrimination Without Indication: Why Dretske Can't Lean on Learning. *Mind & Language, 9*(2), 163–80. https://doi.org/10.1111/j.1468-0017.1994.tb00221.x

Summerfield, Donna and Pat Manfredi (1998). Indeterminancy in Recent Theories of Content. *Minds and Machines*, *8*(2), 181–202. https://doi.org/10.1023/A:1008243329833

Sutton, Richard S. and Andrew G. Barto (2018). *Reinforcement Learning: An Introduction*. MIT press.

Tiffany, Evan (1999). Semantics San Diego Style. *The Journal of Philosophy*, *96*(8), 416–29.

Timberlake, William (2007). Anthropomorphism Revisited. *Comparative Cognition & Behavior Reviews*, 2, 139–44.

Usher, Marius (2001). A Statistical Referential Theory of Content: Using Information Theory to Account for Misrepresentation. *Mind & Language*, *16*(3), 311–34.

Whiten, Andrew (1996). When Does Smart Behaviour-Reading Become Mind-Reading? In P. Carruthers and P. Smith (Eds.), *Theories of Theories of Mind* (277–92). Cambridge University Press.

Wiese, Wanja (2017). What Are the Contents of Representations in Predictive Processing? *Phenomenology and the Cognitive Sciences*, *16*(4), 715–36. https://doi.org/10.1007/s11097-016-9472-0

Williams, Daniel (2018). Predictive Processing and the Representation Wars. *Minds and Machines*, *28*(1), 141–72. https://doi.org/10.1007/s11023-017-9441-6

Wilson, Mark (1982). Predicate Meets Property. *The Philosophical Review*, *91*(4), 549–89.